ICT, Internet and Worker Productivity

Irene Bertschek

Abstract

This article provides a brief overview of the role of information and communication technologies (ICT) as a driver of productivity. In particular, it focuses on the diffusion of computers and the Internet at the workplace and discusses the relationship with wages, the task composition of occupations and labour productivity at the firm level.

Keywords

Complementarities; Computer use; Econometric analysis; Firm-level data; Internet; Labour productivity; Organisational change; Production function; Social networks; Task-based approach; Wage equation

JEL Classifications

D22; D24; J24; J31; O33

Information and communication technologies (ICT) enhance productivity and growth, as shown by various studies at the macroeconomic and microeconomic levels (see for instance Draca et al. 2007 or Kretschmer 2012, for a comprehensive overview). As so-called general purpose technologies (Bresnahan and Trajtenberg 1995), they diffuse throughout the whole economy and enable innovation in adopting firms and sectors (see for example Brynjolfsson and Saunders 2010), leading to higher productivity.

From a technological perspective, we can differentiate three basic stages of ICT: personal computers, the Internet, and more recently, mobile Internet.

When personal computers started to diffuse to workplaces, economists became interested in analysing whether the use of computers makes workers more productive. There are basically two approaches to measuring worker productivity. One approach takes an individual perspective. It is based on the concept of wage functions and assumes that wages reflect individual productivity. The other approach takes a firm-level perspective. It builds on production functions and analyses the relationship between labour productivity and firms' input factors, labour, non-ICT capital and ICT capital, as well as other firm characteristics. Some studies go beyond these main approaches by taking into account job-related tasks, project-based information or regional aspects.

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

Worker Productivity at the Individual Level

Computer Use and Wages

In his seminal paper, Krueger (1993) analyses whether workers who use a computer at work earn higher wages than workers not using a computer at work. The following kind of wage equation is estimated:

$$\ln W_i = \beta_0 + \beta_1 \text{Computer}_i + \beta_2 X_i + u_i$$

where W_i is hourly wage for employee i, Computer_i is a dummy variable taking the value one if employee i uses a computer at work and the value zero otherwise, and X_i represents a vector of employee characteristics, such as education, age and gender. Krueger uses data from the Current Population Survey (CPS) collected in the USA in 1984 and 1989 and from the High School and Beyond Survey for the years 1980, 1982, 1984, 1986. The findings reveal that the wage rate of computer users is about 10–15% higher than that of the non-computer users. In 1984, about 25% of the employees used computers at work, whereas the number has increased to 37% in 1989. Krueger points out that it is not clear from the data and from his analysis whether using computers makes employees more productive and therefore means that they earn higher wages, or whether there are unobserved individual characteristics that correlate with computer use and wages. For instance, high-skilled employees may have abilities that make them earn higher wages and increase the probability of using computers. Owing to the fact that Krueger has only cross-sectional data he cannot control for individual unobserved heterogeneity using fixed effects regression.

DiNardo and Pischke (1997) replied provocatively to Krueger's paper by asking: 'Have pencils changed the wage structure too?'. They replicate Krueger's estimations for the USA but extend the analysis to a further cross-section (1993) of the Current Population Survey. Furthermore, they use German employee-level data from the Qualification and Career Survey for the time periods 1979, 1985/86 and 1991/92 and compare the results to those found for the USA. In addition to looking at computers at work, the authors also consider the effect of other working tools such as calculators, telephones, writing tools like pens or pencils, and sitting on the job. They call these tools 'white-collar tools', since they are more probably used by white-collar than by blue-collar workers. The results for Germany, with respect to the wage premium for computer use, confirm the results found for the USA. However, similar wage differentials are found for pens and pencils, calculators, telephones and working while sitting. The authors draw the conclusion that the wage differential found for computer use cannot reflect true returns to computer use or skills, since otherwise no similar effects would have been found for the other white-collar tools. Similar results are found by Borghans and Ter Weel (2004) for British employee data collected in 1997. Computer use only at the advanced level is related to wage premiums, whereas mathematics and writing skills show significant wage premiums.

The study by Entorf et al. (1999) has the advantage over previous studies that it relies on panel linked employer–employee data for 1991 to 1993. This allows individual fixed effects that control for unobserved heterogeneity across employees to be taken into account. Moreover, the authors can observe what happens if an employee starts using a computer at work. The wage differential observed in the cross-sectional French data is more or less the same as in the USA and lies between 15 and 20%. Panel regressions, however, show that this wage differential decreases to only up to 2%, a result that confirms evidence found before for the 1980s (Entorf and Kramarz 1997). Moreover, employees were already better paid before they started using computers. This result implies that firms allocate computers to selected workers and these workers seem to have unobserved skills that are complementary with computer use. According to the French dataset, this seems to hold particularly for low-skilled workers. Wage premium estimates are summarised in Table 1.

The Task-Based Approach

In order to obtain deeper insights into the unobserved characteristics that are complementary

ICT, Internet and Worker Productivity, Table 1 Overview of wage premium estimates

Authors	Data	Estimated wage premium
Krueger (1993)	US Data: Current Population Survey, 1984, 1989; High School and Beyond Survey, 1980, 1982, 1984, 1986	Between 10% and 15% for computer use
DiNardo and Pischke (1997)	US Data: Current Population Survey, 1984, 1989, 1993; West German Data: Qualification and Career Survey, 1979, 1985–1986, 1991–1992	For the USA 1989: 19% For West Germany 1991: 17% for computer use Similar effects for pencils and other white-collar tools
Entorf and Kramarz (1997)	French Labour Force Survey 1985–1987 and firm-level information	Cross-section analysis: 16% for computer-related new technologies, decomposed in 6% for workers with zero experience and 2% for each year of experience for the first years
		Longitudinal individual data: effect not related to experience disappears 1% for each year of experience
Entorf et al. (1999)	French Labour Force Survey 1991–1993 and firm-level information	Cross-section analysis: 15–20% Panel data: 2%
Borghans and Ter Weel (2004)	Skills Survey of the Employed British Workforce 1997	Wage returns only if computers are used at the advanced level (e.g. programming) Positive and significant wage premiums for mathematical and writing skills

with computer use, Autor et al. (2003) suggested a so-called task-based approach. This approach assumes that work consists of a series of routine and non-routine tasks. While manual and cognitive routine tasks can be performed and thus substituted by a computer, non-routine tasks cannot. Analytical and interactive non-routine cognitive tasks are, by contrast, supported (i.e. complemented) by computers. For instance, doing research or advising customers are non-routine cognitive tasks that can be better performed using a computer. By contrast, bookkeeping or controlling machines can be performed by computers (see Spitz-Oener 2006, p. 243, for a classification of tasks). Autor et al. (2003) and Spitz-Oener (2006) have shown for the USA and Germany, respectively, that the diffusion of computers goes hand in hand with a shift in the content of work from manual and cognitive routine tasks towards non-routine cognitive tasks. This shift implies an increase in the demand for skilled employees (in line with the hypothesis that technological change is skill-biased), leading to increased wages for these skills. Another, more direct, channel for how computers affect wages is that employees become increasingly productive when complementing their tasks with computer use.

This latter aspect corresponds to the complementarities between computer use and organisational change found at the firm level by Bresnahan et al. (2002) (see next section). For the empirical analysis of the task-based approach, task compositions within occupations are calculated for each employee i (see Spitz-Oener 2006, p. 242 for the following definition):

$$\text{Task}_{ijt} = \frac{\text{number of activities in category } j \text{ performed by } i \text{ at time } t}{\text{total number of activities in category } j \text{ at time } t}$$

where j represents the tasks, i.e. $j = 1$ (nonroutine analytic tasks), $j = 2$ (nonroutine interactive tasks), $j = 3$ (routine cognitive tasks), $j = 4$ (routine manual tasks) and $j = 5$ (nonroutine manual tasks), and t reflects the cross-section for which data is available. According to the example

given by Spitz-Oener, if employee *i* performs two out of four analytical activities, his or her analytical task measure is 50. Based on these task measures the change in the shares of tasks within occupations over time can be calculated, showing which of the tasks have become more or less important.

Spitz-Oener (2008) extends her previous analysis conducted in 2006 to take up the issue raised by DiNardo and Pischke (1997), i.e. that there is also an effect of pencil use on wages. She shows for West German employee data, again from the Qualification and Career Survey in 1998/99, that wage premiums are observed for employees with skills that are complementary with computer use. In contrast with the study by DiNardo and Pischke (1997), no similar effects are found for the use of pencils. This result underpins what has been found before: computer use has shifted the task composition of occupations towards analytical and interactive tasks and away from routine cognitive and manual tasks. While computers complement the first, they tend to substitute for the latter.

Worker Productivity at the Firm Level

ICT and Labour Productivity

Taking a firm-level perspective, the relationship between labour productivity and ICT can be captured by a production function approach. Output Q is related to the input factors of labour, non-ICT capital and ICT capital. Although sometimes materials explicitly are taken into account, we do not consider them here. If we assume a Cobb-Douglas form of the production function, the function for firm i looks as follows:

$$Q_i = AL_i^{\beta_1} K_i^{\beta_2} C_i^{\beta_3}$$

where Q is output, L is labour, K is capital, C is ICT capital and A represents a technology or efficiency parameter. The parameters β_1, β_2 and β_3 represent the output elasticities of the respective input factors. Taking logarithms, setting $\ln A = \beta_0$ and adding an error term u_i leads to the following equation:

$$\ln Q_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i + \beta_3 \ln C_i + u_i$$

Subtracting $\ln L$ from both sides results in

$$\ln\left(\frac{Q_i}{L_i}\right) = \beta_0 + (\beta_1 - 1)\ln L_i + \beta_2 \ln K_i + \beta_3 \ln C_i + u_i$$

where $\ln\left(\frac{Q_i}{L_i}\right)$ represents labour productivity, i.e. output per worker. This equation can be estimated by econometric methods using firm-level data. If panel data are available, an index t is added. Panel data usually allow taking account of firm-specific fixed effects and thus unobserved heterogeneity across firms. Depending on the available data, labour productivity is measured by sales per employee, sales per hour worked, value added per employee or value added per hour worked. ICT capital often is not observable in firm-level data sets. In this case, it may be approximated by ICT investment or by the percentage of employees working with computers. If panel data is available and information about ICT investment, then ICT capital stocks can be calculated according to the so-called perpetual inventory method (see for example Bloom et al. 2012, or Hempell 2005). Some studies, instead of using measures of ICT capital, analyse the relationship of labour productivity with specific ICT applications such as B2B e-commerce that are measured by dummy variables (see for example Bertschek et al. 2006).

There is meanwhile a large number of firm-level studies analysing the role of ICT for labour productivity. Draca et al. (2007) provide a comprehensive overview of the studies published between 1996 and 2005 and summarise the main results. One main finding of most studies is that labour productivity is positively and significantly related with ICT. The average of the estimated coefficients of ICT is about 5% to 6% and has increased over time (see Kretschmer 2012). This relationship, however, might be heterogeneous with respect to firms and industries, i.e. some firms or industries are more successful in employing ICT than others.

This firm-specific heterogeneity in reaping the potential of ICT might be due to differences in

complementary investment in organisational capital and human capital across firms, an argument put forward for instance in Bresnahan et al. (2002) and underpinned in several further studies. The relationship between productivity and ICT is stronger if investment in ICT is supported by investment in organisational capital (for the particular role of organisational capital see for example Black and Lynch 2001, for the USA; Bertschek and Kaiser 2004, for Germany). Since ICT lowers the cost of communication, employees can communicate and exchange information more efficiently. Thus, working in teams and with a low number of hierarchies becomes more feasible and organisational structures may become more decentralised and flexible. Moreover, communicating and coordinating with customers and suppliers become easier and costs may decrease. Investment in human capital is considered to be complementary with ICT since the implementation of a new ICT system or application in a firm often requires that firms train their employees in order to be able to work with these new technologies or applications. Recent evidence on the relationship between ICT, organisational capital and human capital and its productivity-enhancing effect is presented by Bloom et al. (2012) who find that US multinationals located in Europe obtain higher productivity effects from using ICT than their non-US counterparts due to better people management practices. Bartel et al. (2007) analyse specific ICT applications in valve-producing plants, and Aral et al. (2012) present econometric evidence on a three-way complementarity between firms' adoption of software for human capital management, performance pay and firms' practice of human resource analytics (including worker monitoring, performance feedback, the integration of workforce support data, and talent management). Hall et al. (2012) consider investment in ICT and in research and development (R&D) as potential sources of innovation which in turn may enhance labour productivity. They use four cross sections of Italian manufacturing firms covering the period 1995–2006. The econometric results show that R&D and ICT contribute directly to labour productivity but also indirectly through enabling innovation.

One big issue empirically working economists are faced with is endogeneity. It is *a priori* not evident whether investment in ICT increases labour productivity or whether productivity growth implies more investment in ICT. Depending on the available datasets, the studies are more or less able to tackle this issue.

Looking at the Internet as a specific ICT, there is not so much empirical work yet as exists for ICT in general. The following section will summarise some of the empirical results.

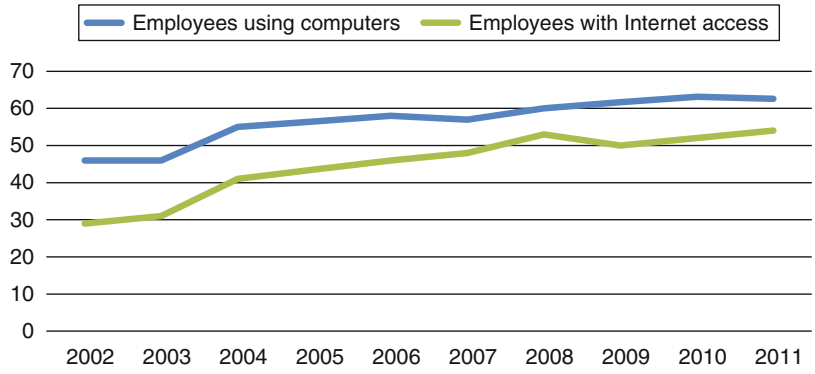
Internet and Labour Productivity

The Internet started to diffuse to workplaces later than computers. If a firm connects to the Internet, it does not necessarily mean that all employees have Internet access, but this might rather be restricted to a certain group of persons such as the chief executive officers or the administration staff. Also, Internet use as well as computer use varies considerably between manufacturing and services industries.

Figure 1 shows the diffusion of computers and Internet access in German firms. The percentage of employees using a computer at least once per week at work has increased from 46% in 2002 to 63% in 2011. In the same period, the percentage of employees with connection to the Internet has increased from 29% in 2003 to 54% in 2011.

What does the Internet add to only having a computer? It allows access to information and connects employees with each other, facilitating the search and exchange of information. Moreover, Internet technologies or web-based applications such as wikis or collaboration platforms facilitate processing of information, documentation and cooperation.

For the case of New Zealand, Grimes et al. (2012) find based on a firm-level cross section collected in 2006 that firms using broadband Internet have a 7 to 10% higher labour productivity. By contrast, for the early phase of broadband diffusion in Germany, 2000 to 2002, Bertschek et al. (2011) find positive and significant effects of broadband on firms' innovation activity but not on labour productivity. Polder et al. (2010) analyse the role of ICT and R&D for innovation success. Their estimations are



ICT, Internet and Worker Productivity, Fig. 1 Percentage of employees using computers and Internet in German firms, 2002–2011. In 2008, NACE code classification has changed

based on three data waves of Dutch firms. ICT is measured as investment in ICT per employee. Additionally, they use a measure for Internet use (the percentage of employees having access to broadband Internet), and include e-commerce as a specific ICT application. The results show that broadband Internet is particularly important for service firms, where broadband is positively related to product and process innovation as well as to organisational innovation. By contrast, in the manufacturing sector, broadband is significant only for product and organisational innovation. For process innovation, it is e-commerce that plays a significant role. These results support the hypothesis of complementarity between ICT and innovation or organisational change. Moreover, they show that it also depends on what firms concretely do with their ICT or with their Internet access to enable innovation. This latter issue is taken up in a recent paper by Colombo et al. (2012). The authors show for a sample of small Italian firms that it is not the broadband connection itself that makes firms more productive. It depends rather on the kind of application as well as on complementary organisational and strategic changes whether or not firms profit with respect to their productivity.

Internet and Wages: A Regional Perspective

Studies looking at Internet and wages are still scarce. Forman et al. (2011) take a regional perspective. Their initial hypothesis is that Internet lowers the cost for economic engagement also in geographically isolated regions. Thus, Internet should have effects on the performance of firms and employees also in regions whose performance was comparably low before the diffusion of the Internet. The study does not look at broadband Internet itself but at business investment in advanced Internet technologies. These comprise investment in enterprise resource planning (ERP), customer service, education, extranet, publications, purchasing and technical support. The time span of the analysis is from 1995 to 2000, a time period when Internet had just started to diffuse more broadly and when there was still a lot of variation in the use of broadband Internet or Internet-based applications with respect to firms, individuals and regions. The authors use data from different sources on firms with more than 100 employees as well as county-level data.

The estimations show that although advanced Internet applications diffused widely in the USA from 1995 to 2000, the economic benefits in terms of wage growth were concentrated in a few well-performing counties only. More precisely, only 6% of US counties profited from investment in Internet technologies in terms of wage growth. This wage growth amounted to 28% from 1995 to 2000, whereas the average growth over all counties was 20%. These counties, however, had a better performance already before 1995, i.e. they were characterised by relatively high income, large population, high skills and high IT intensity. The results of the study thus do not support the initial hypothesis

that the Internet contributes to economic regional inclusion, but rather imply that the Internet aggravates regional wage inequality.

Including Social Network Data

In order to analyse the relationship between multitasking, knowledge networks and productivity, the approach by Aral et al. (2012) goes beyond the firm level and the individual level. The authors focus on only one firm, a mid-size executive recruiting firm. They use detailed data on employees' characteristics, on their project output and team membership for projects, and on email messages sent and received by these employees, i.e. on the workers' digital network. A recruiting firm offers services, and for services, measuring output, input and productivity is harder than in manufacturing firms. The authors of the study have accounting records for all projects covering the period 2001 to 2005, including the number of projects completed and the revenue generated by individual recruiters. They measure output as the number of projects completed per month, i.e. the number of days a recruiter works on the project per month divided by the total number of days for which the project runs. Completing a project means that the recruiter has found an appropriate candidate for the client and the candidate has signed a contract. Output is set into relation with the heterogeneity of multitasking measured as the number of projects recruiters work on per month and the heterogeneity of recruiters' contacts resulting from the work on prior projects as well as from the number of email contacts.

There are several interesting findings from this analysis: Recruiters' output is increasing with multitasking, but only up to a certain threshold. A further increase of multitasking then implies diminishing rates of return. Although heterogeneity of contacts is negatively related with output, it complements task heterogeneity. Having access to heterogeneous information via email makes multitasking recruiters more productive. This result again supports the hypothesis of complementarity between workplace organisation and IT as well as the complementarity between specific tasks and IT.

Current Technological Trends

While computers allow for digitisation, and the Internet for connectedness, the mobile Internet, which has started to diffuse only recently, additionally offers the possibility to work at any time from any place. For example, in Germany, on average 25% of employees have broadband access via a mobile device such as a smartphone or a tablet (Statistisches Bundesamt 2011, p. 21). This development is supported by so-called cloud computing – the concentration of computing capacity, data and software in data centres that employees can connect to from anywhere. There are so far no empirical econometric studies based on large-scale data analysing whether mobile Internet adds to worker productivity additionally to computers and Internet. We can imagine what might happen if working environments get more and more flexible and independent from time and space. On the one hand, this technological opportunity, by decreasing information and communication costs, supports further decentralisation of work as suggested in Bresnahan et al. (2002). On the other hand, these flexible working environments require a high degree of self-reliance and coordination. Very probably, new studies will soon give insights into the net effects of these new technological advances.

See Also

- ▶ Internet, Economics of the
- ▶ Production Functions
- ▶ Social Networks, Economic Relevance of

Bibliography

Aral, S., E. Brynjolfsson, and M. van Alstyne. 2012. *Information, technology and information worker productivity.* Working paper. Available at: http://ssrn.com/abstract=942310

Aral, S., E. Brynjolfsson, and L. Wu. 2012b. Three-way complementarities: Performance pay, human resource analytics, and information technology. *Management Science* 58(5): 913–931. doi:10.1287/mnsc.1110.1460.

Autor, D.H., F. Levy, and R.J. Murnane. 2003. The skill content of recent technological change: An empirical

exploration. *Quarterly Journal of Economics* 118(4): 1279–1333.
Bartel, A., C. Ichniowski, and K. Shaw. 2007. How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. *Quarterly Journal of Economics* 122(4): 1721–1758. doi:10.1162/qjec.2007.122.4.1721.
Bertschek, I., and U. Kaiser. 2004. Productivity effects of organizational change: Microeconometric evidence. *Management Science* 50(3): 394–404.
Bertschek, I., H. Fryges, and U. Kaiser. 2006. B2B or not to be: Does B2B e-commerce increase labour productivity? *International Journal of the Economics of Business* 13(3): 387–405.
Bertschek, I., D. Cerquera, and G.J. Klein. 2011. More bits – More bucks? Measuring the impact of broadband internet on firm performance, ZEW Discussion Papers, No. 11–032.
Black, S., and L. Lynch. 2001. How to compete: The impact of workplace practices and information technology on productivity. *The Review of Economics and Statistics* 83(3): 434–445.
Bloom, N., R. Sadun, and J. Van Reenen. 2012. Americans do IT better: US multinationals and the productivity miracle. *American Economic Review* 102(1): 167–201.
Borghans, L., and B. ter Weel. 2004. Are computer skills the new basic skills? The returns to computer, writing, and math skills in Britain. *Labour Economics* 11(1): 85–98.
Bresnahan, T.F., and M. Trajtenberg. 1995. General purpose technologies 'engines of growth'? *Journal of Econometrics* 65(1): 83–108.
Bresnahan, T.F., E. Brynjolfsson, and L.M. Hitt. 2002. Information technology, workplace organization and the demand for skilled labour: Firm-level evidence. *Quarterly Journal of Economics* 117(1): 339–376.
Brynjolfsson, E., and A. Saunders. 2010. *Wired for innovation: How information technology is reshaping the economy*. Cambridge, MA: MIT Press.
Colombo, M.G., A. Croce, and L. Grili. 2012. *ICT services and small businesses' productivity gains: An analysis of the use of broadband Internet technology*. Working paper. Milano.
DiNardo, J.E., and J.-S. Pischke. 1997. The returns to computer use revisited: Have pencils changed the wage structure too? *Quarterly Journal of Economics* 112(1): 291–303.
Draca, M., R. Sadun, and J. Van Reenen. 2007. Productivity and ICTs: A review of the evidence. In *The oxford handbook of information and communication technologies*, ed. R. Mansell, C. Avgerou, D. Quah, and R. Silverstone, 100–147. New York: Oxford University Press.
Entorf, H., and F. Kramarz. 1997. Does unmeasured ability explain the higher wages of new technology workers? *European Economic Review* 41(8): 1489–1509.
Entorf, H., M. Gollac, and F. Kramarz. 1999. New technologies, wages, and worker selection. *Journal of Labor Economics* 17(3): 464–491.
Forman, C., A. Goldfarb, and S. Greenstein. 2011. The Internet and local wages: A puzzle. *American Economic Review* 102(1): 556–575.
Grimes, A., C. Ren, and P. Stevens. 2012. The need for speed: Impacts of Internet connectivity on firm productivity. *Journal of Productivity Analysis* 37(2): 187–201.
Hall, B.H., F. Lotti, and J. Mairesse. 2012. Evidence on the impact of ICT investment on innovation and productivity in Italian firms. *Economics of Innovation and New Technology*. doi:10.1080/10438599.2012.708134.
Hempell, T. 2005. What's spurious? What's real? Measuring the productivity impacts of ICT at the firm level. *Empirical Economics* 30(2): 427–464.
Kretschmer, T. 2012. Information and communication technologies and productivity growth: A survey of the literature. *OECD Digital Economy Papers*, No. 195. OECD Publishing. doi:10.1787/5k9bh3jllgs7-en.
Krueger, A.B.. 1993. How computers have changed the wage structure: Evidence from microdata, 1984–1989. *Quarterly Journal of Economics* 108(1): 33–60.
Polder, M., G. van Leeuwen, P. Mohnen, and W. Raymond. 2010. *Product, process and organizational innovation: Drivers, complementarity and productivity effects*. UNU-MERIT Working Paper (2010–035).
Spitz-Oener, A. 2006. Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *Journal of Labor Economics* 24(2): 235–270.
Spitz-Oener, A. 2008. Returns to pencil use revisited. *Industrial and Labor Relations Review* 61(4): 502–517.
Statistisches Bundesamt. 2011. Unternehmen und Arbeitsstätten. Nutzung von Informations- und Kommunikationstechnologien in Unternehmen. Wiesbaden.

Ideal Indexes

Kazuo Sato

Among many index numbers, the two most favoured because of algebraic simplicity and ease of computation are those advocated by E. Laspeyres in 1864 and by H. Paasche in 1874. There are n commodities, indexed from 1 to n. At time point t, the price vector is $p_t = \{p_{1t}, \ldots, p_{nt}\}$ and the quantity vector $q_t = \{q_{1t}, \ldots, q_{nt}\}$. $p_s q_t$ denotes $\sum_{i=1}^{n} p_{is} q_{it}$. Let P_{st} and Q_{st} be the price and quantity indexes from time s to t. Then, these two indexes are

$$\text{Laspeyres} P_{st}^{L} = p_t q_s / p_s q_s, \quad Q_{st}^{L} = p_s q_t / p_s q_s$$
$$\text{Paasche} P_{st}^{P} = p_t q_t / p_s q_t, \quad Q_{st}^{P} = p_t q_t / p_t q_s$$

There are several desirable properties that an index ought to satisfy (Samuelson and Swamy 1974; Allen 1975, pp. 40–47). Three basic tests (stated for the price index) which any reasonable index must meet are:

1. Identity test: $P_{tt} = 1$.
2. Proportionality test: $P_{st'} = kP_{st}$, when $p_{it'} = kp_{it}$, $q_{it'} = q_{it}$ for all i.
3. Dimensional test: changes of units do not affect the index value. The next three are not always satisfied:
4. Time-reversal test: $P_{st}P_{ts} = 1$.
5. Circular test: $P_{rs}P_{st} = P_{rt}$.
6. Factor-reversal test: $P_{st}Q_{st} = E_{st}$, where $E_{st} = p_t q_t / p_s q_s$ is the expenditure index and P and Q are matching indexes in the sense that they share a common form except that p and q are interchanged between them.

Irving Fisher (1922), who most energetically pursued the topic of index numbers, emphasized the factor-reversal test and regarded PQ/E (where P and Q are matching indexes) as the bias of an index. Very few indexes satisfy (6). For the Laspeyres and Paasche indexes, the following identities are seen to hold:

$$P_{st}^{L} Q_{st}^{P} = P_{st}^{P} Q_{st}^{L} = E_{st},$$

i.e., the Laspeyres and Paasche indexes are 'factor antitheses'. Then, their geometric averages

$$P_{st}^{F} = \sqrt{P_{st}^{L} P_{st}^{P}}, \; Q_{st}^{F} = \sqrt{Q_{st}^{L} Q_{st}^{P}}$$

satisfy (6). Fisher regarded this index to be the best or 'ideal' among 134 indexes he compared. This index has been known as Fisher's ideal index even though he was not the only one who discussed this index at the time.

A log-change index has also been popular. It is given the form

$$\ln P_{st} = \sum_i s_i (\ln p_{it} - \ln p_{is}),$$

$$\ln Q_{st} = \sum_i s_i (\ln q_{it} - \ln q_{is})$$

where $s_i \geq 0$, $\sum s_i - 1$. Expenditure shares are used for weights. Let w_{it} be the share of good i in total expenditure at time t. Loglinear analogues of the Laspeyres, Paasche, and Fisher indexes are obtained by setting (i) $s_i = w_{is}$, (ii) $s_i = w_{it}$, and (iii) $s_i = \frac{1}{2}(w_{is} + w_{it})$. The last one, which is attributed to Törnqvist, does not satisfy the factor-reversal test.

Log-change indexes may be considered as discrete approximations to the continuous Divisia index obtained by integrating

$$d \ln P = \sum_i w_i d \ln p_i, \quad d \ln Q = \sum_i w_i d \ln q_i$$

from s to t.

Suppose that (p, q) represents the behaviour of a consumer maximizing utility. Assume that the consumer's utility is represented by a preference function of a certain homogeneous form. It can then be shown that the Divisia index also assumes a certain form. This index is said to be 'exact' with the preference function (Diewert 1976). (The Laspeyres is exact with a linear utility function, the Paasche with a Leontief-type utility function, and the Törnqvist with a translog utility function.)

The correspondence between a preference function and an index can be given the following heuristic argument: The preference ordering can be represented either in a direct form $[U(q)]$ or in an indirect form $[V(E/p)]$. Interpreting the quantity index as a constant-utility index, we have $Q_{st} = U(q_t)/U(q_s)$. By the same token, the price index is associated with the indirect utility function so that $P_{st} = V(E/p_t)/V(E/p_s)$. When U and V are alternative representations of a preference function, they form a dual pair. P and Q which are exact with them are factor antitheses, namely, $P_{st}Q_{st} = E_{st}$. As P and Q are not in general

matching indexes, they do not meet the factor reversal test.

When the duel pair, U and V, share a common functional form, they are called 'self-dual' (Houthakker 1965). It then follows that Q and P which are exact with the dual pair must also share a common form, i.e., they are matching indexes. Thus, an important proposition holds: there are as many ideal index numbers as there are self-dual preference functions as they are equivalent to each other. There are only three known self-dual preference functions: (-a) Cobb–Douglas, (b) quadratic, and (c) constant-elasticity-of substitution (CES). Ideal indexes which correspond to these are as follows:

(a) A log-change index with fixed weights. The weights are exponents of the Cobb–Douglas. Since expenditure shares do not remain constant over time, this index violates reality.
(b) Fisher's ideal index. This correspondence was noted by Konüs and Byushgens already in the 1920s (Afriat 1977).
(c) A log-change index with variable weights where s_i is given by $(w_{it}–w_{is})/(\ln w_{it}–\ln w_{is})$, divided by its sum over i. Though complicated in form, these weights are seen to be in the nature of geometric averages. This index was discovered independently by Sato (1976) and Vartia (1976).

No other self-dual preferences not ideal indexes have been discovered since.

See Also

► Index Numbers

Bibliography

Afriat, S.N. 1977. *The price index*. London: Cambridge University Press.
Allen, R.G.D. 1975. *Index numbers in theory and practice*. Chicago: Aldine.
Diewert, W.E. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4: 115–145.
Fisher, I. 1922. *The making of index numbers*. Boston: Houghton Mifflin.
Houthakker, H.S. 1965. A note on self-dual preferences. *Econometrica* 33: 797–801.
Samuelson, P.A., and S. Swamy. 1974. Invariant economic index numbers and canonical duality: Survey and synthesis. *American Economic Review* 64: 566–593.
Sato, K. 1976. The ideal log-change index number. *Review of Economics and Statistics* 58: 223–228.
Vartia, Y.O. 1976. Ideal log-change index numbers. *Scandinavian Journal of Statistics* 3: 121–126.

Ideal Output

J. V. de Graaff

Abstract

Pigou's notion of 'the ideal output' as 'the output in any industry which maximizes the national dividend, and, apart from the differences in the marginal utility of money to different people, also maximises satisfaction' has long been eclipsed by the 'general optimum of production and exchange', in which the welfare of each member of the community is maximized in turn, subject to certain constraints – even though the more modern theory, despite its advantages, does not necessarily reach any substantially different conclusions. But ideal output theory is by now no more than an episode in the history of economic thought.

Keywords

Barone, E.; Competitive equilibrium; External economies; General optimum of production and exchange; Ideal output; Imperfect competition; Marginal equivalence; Marginal social product; Monopoly; Pareto, V.; Pigou, A. C

JEL Classifications

D6

Pigou, writing in *The Economics of Welfare,* calls 'the output in any industry which maximizes the national dividend, and, apart from the differences in the marginal utility of money to different

people, also maximises satisfaction, the ideal output'. He goes on to argue that 'this output is attained – the possibility of multiple maximum positions being ignored – when the value of the marginal social net product of each sort of resource invested in the industry under review is equal to the value of the marginal social net product of resources in general'. And, finally, it 'will be that output which makes the demand price of the output equal to the money value of the resources engaged in producing a marginal unit of output' (1932, pp. 802, 803).

The line of argument that comes through so clearly in these quotations can be traced back to Pigou's earlier *Wealth and Welfare* (1912) and indeed to Marshall; but since the 1930s it has been overtaken by the development of a more powerful strand of analysis that stems from Pareto (1897) and Barone (1908) and has culminated in the theory of the general optimum of production and exchange. In it one maximizes in turn the welfare of each member of the community, subject to the constraint of the social production function and to holding on each occasion the welfare of each other member constant. The resulting first-order conditions include the marginal equivalences enumerated in the theory of ideal output (Graaff 1957). Any modern discussion of the theory must therefore be set against the background of the one that has incorporated and replaced it.

The more modern theory has the virtues of elegance, simplicity and generality. It embraces exchange as well as production. It deals with commodities and firms (or event plants) instead of industries. It does not need the doctrine of maximum satisfaction, or any assumption about interpersonal comparisons of utility. But at the end of the day it does not reach any substantial conclusion that the theory of ideal output, correctly employed, would not itself have reached.

The problem, especially in the early development of the theory, was that it was not all that easy to apply it correctly. It was not originally recognized that (at least in a closed economy) the correct way to reckon the value of a marginal social net product is at *constant* prices. The same remark applies to the calculation of marginal social cost. If higher prices have to be offered to factors of production to attract them to an industry undergoing expansion, the element of the cost of the expansion caused by the higher prices represents a transfer payment to the factors (in the form of a rent or quasi-rent), not a cost to society. The cost to society is the value of the output sacrificed when the factors are withdrawn from their previous use. That value was reckoned at the original prices of the factors. Those prices must therefore be used in reckoning their cost to society in their new use.

Clarification of this issue was the result of a famous debate of the 1920s – much of it reprinted in *Readings in Price Theory* (Stigler and Boulding 1953) – on the desirability of taxing industries subject to diminishing returns, and paying bounties to those subject to increasing returns, a result to which the theory of ideal output at one stage seemed to point. As competitive conditions were meant to be prevailing, the industries enjoying increasing returns had to be assumed to comprise firms whose unit costs were falling because of *external* economies; and as external economies were themselves recognized as possible reasons for a divergence between private and social net products, the opportunities for getting muddled were legion. It is to the credit of the participants – among them D.H. Robertson, G.F. Shove, F.H. Knight and J. Viner – that these dangers were largely avoided.

Much of the motivation for the theory of ideal output seems to have been a desire to see when competitive output was ideal, and when interference in a competitive economy would be justified. Today we ask, rather more formally (cf. Debreu 1959), when a competitive equilibrium would also be a general optimum. The answer, very briefly, is when the technology is convex, there are no external effects in production or consumption, no public goods and no foreign trade.

Apart from the fact that the existence of public goods was glossed over, ideal output theory would not have given a very different answer. The importance of the foreign trade exception was recognized. (The marginal social cost of importing goods subject to rising supply price is higher than the marginal private cost. The rents that accrue to *foreigners* are not mere transfers

within the domestic community, but a part of social cost). Divergences between private and social costs due to external economies and diseconomies in production, and between private and social benefits due to external economies and diseconomies in consumption, were fully discussed. The counterpart of the modern insistence on a convex technology was the painstaking treatment of increasing returns. The conditions under which competitive output would approach the ideal were pretty clearly defined.

Pigou also discussed the deviation from the ideal of the outputs of discriminating monopolists. (Not surprisingly, they fell short.) R.F. Kahn (1935) extended the analysis to imperfect competition. He argued that (taking diseconomies as negative economies) all industries could be arranged in descending order on a scale according to the extent of the external economies they generated and the degree of monopoly (measured by the gap between price and marginal cost) they enjoyed and that at a certain point on the scale there would be an average industry. Above this point all should expand to produce ideal outputs; below it all should contract. Adjustment could be achieved by a set of taxes and bounties. When all industries had expanded or contracted to conform to the average degree of monopoly and the average capacity to create external economies, their marginal social products would diverge from their marginal private products to the same extent and ideal output would be attained.

Note that this treatment avoids the error of making 'piecemeal' recommendations of the sort so often found in partial analysis. All industries must move to the average. It may not help if one or two do. That may just increase the gap between those that conform and those that do not. (In technical terms, the first-order conditions for a maximum must be satisfied simultaneously.)

In this sense Kahn's treatment is very general. In another it is not general enough. Proportionality of marginal products is not sufficient. For a full optimum, equality is essential (Lerner 1944, ch. 9). This may require an adjustment in the number of hours worked, and an expansion or contraction in the level of output as a whole.

The view that suitable corrective taxes and bounties can and should be used to bring marginal private products into line with marginal social products, when they diverge, was once very popular. On the whole it has weathered less well than ideal output theory itself, although the latter is by now no more than an episode in the history of economic thought.

See Also

► Pareto Efficiency

Bibliography

Barone, E. 1908. II MinisteriodellaProduzionenellostato-collectivista. *GiornaledegliEconomisti.* Trans. *Collectivist economic planning,* ed. F.A. Hayek. London: Routledge, 1935.

de Graaff, J.V. 1957. *Theoretical welfare economics*. Cambridge: Cambridge University Press.

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

Kahn, R.F. 1935. Some notes on ideal output. *Economic Journal* 45: 1–35.

Lerner, A.P. 1944. *The economics of control*. New York: Macmillan.

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

Pigou, A.C. 1912. *Wealth and welfare*. London: Macmillan.

Pigou, A.C. 1932. *The economics of welfare,* 4th ed. London: Macmillan.

Stigler, G.J., and K.E. Boulding, eds. 1953. *Readings in price theory*. London: Allen &Unwin.

ideal type

David Beetham

This is the term used by Max Weber to describe the distinctive concepts and models developed by economic and social theorists, and employed in the activity of empirical analysis. The term also defines

the characteristic method which Weber saw as distinctive of the social sciences. Social life is infinitely complex and can never be exhaustively described or explained. In order to make sense of it, the social scientist uses artificially pure concepts, e.g. 'natural economy', 'handicraft', 'capitalism', which are intellectual constructs involving a high degree of abstraction from the actual world. They comprise the most typical elements which have been isolated from a historically repeated pattern of action, relationship or institution, as seen from a partial point of view (economic, political, etc.), and combined into an internally consistent and inherently intelligible unity. With the help of such constructs the social scientist is able to characterize a particular object of study, and make its complexity intelligible according to its degree of conformity to the stipulations of the relevant concept or model. Often a particular social complex will require a combination of such concepts for its elucidation, as for example the class structure of a given society can be understood as a combination of the analytically separable elements of property ownership ('class'), social esteem ('status') and authority position ('power'). Ideal types have nothing to do with ideals (though there can be ideal-type of ideals, e.g. 'individualism') and could perhaps less confusingly be called 'pure types'.

Weber's characterization of the ideal-type method is best understood in the context of the 'Methodenstreit' between the historical and theoretical schools of German political economy. He developed it to rebut what he saw as a mistaken understanding of theory on the part of certain members of the historical school. In their view economic theory should involve the quest for universal laws of a natural-scientific kind, arrived at inductively on the basis of exhaustive empirical studies of economic phenomena. They saw the work of the historical school as the necessary preliminary stage to the discovery of such laws. Measured against this conception, the theoretical work of Carl Menger and the marginal utility school was judged to be excessively abstract, one-sided in its assumptions about human nature, and above all premature.

Weber's ideal-type method provided a critique of this 'scientistic' understanding of economic and social theory. The focus of interest of the social scientist, he argued, lay in the historically specific, not the most general, aspects of phenomena. The latter were both the most banal and the least useful for explanatory purposes. The distinctive method of social-scientific abstraction involved not a quest for universal laws, but a process of isolating what was most typical and essential to a pattern of action or social relation, and rendering it intelligible as an internally coherent whole. It was by the same method of abstraction that the typical historical preconditions and consequences of a given social institution were to be elucidated. Weber argued that this was in practice the method adopted by economic theoreticians, Marxists and marginalists alike, though they did not always recognize its implications. The error of the former was to treat their theoretical deductions about the typical consequences of capitalist competition as actual historical tendencies or laws, in advance of any empirical confirmation. The error of the latter was their failure to recognize the actual historical preconditions for the rigorous calculation and maximization of economic interests, which made their theoretical models historically specific rather than applicable to all times and places (Weber 1903; 1904a).

In this manner Weber's account of the ideal-type method offered a resolution of the controversy between the historical and theoretical schools. On the one hand it demonstrated the historical specificity of even the most abstract theorizing. On the other hand it revealed the irreducibly theoretical character of the concepts used in historical economics, which was anything but a merely descriptive activity on whose successful 'completion' the construction of theory was itself supposedly dependent. Properly understood, the respective emphases of theory and history were mutually complementary, a conjunction which Weber's own work such as *The Protestant Ethic and the Spirit of Capitalism* (Weber 1904b) or the more theoretical formulations of *Economy and Society* (Weber 1921) amply demonstrated.

Subsequent discussion of the ideal-type method has taken place within sociology and political science, rather than within the discipline

of economics, which provided its original intellectual location. Most social scientists would accept the necessity for typological construction, but disagree over both its manner and the criteria for its assessment. Weber's approach has been criticized for its inherent subjectivity, in two quite different senses. First, his method of 'Verstehen' or 'understanding', which is necessary for assessing the internal coherence of idealtype constructs, has been seen as unavoidably arbitrary. To this it can be simply replied that the criteria for the intelligibility of social action are interpersonal, not private, despite the obvious difficulties in respect of alien cultures.

Secondly, following the neo-Kantianism of Heinrich Rickert, Weber argued that the objects of study and hence the concepts used in the social sciences are determined according to their 'value-relevance', i.e. their significance for our values. Unlike Rickert, however, he did not believe that these value standpoints could be objectively grounded in human reason. Some commentators have therefore concluded that it is impossible to rescue Weberian concept formation from the subjectivity of the investigator's own values. Such a conclusion overlooks Weber's insistence that ideal-type constructs must satisfy the criterion of explanatory power as well as of significance, and thus 'be valid for all who seek the truth'. The ultimate test for idealtype construction must be an objective one: its fruitfulness in identifying and resolving explanatory problems.

See Also

► Weber, Max (1864–1920)

Bibliography

Weber, M. 1903. In *Roscher and Knies: The logical problems of historical economics*, ed. G. Oakes. New York: Free Press, 1975.

Weber, M. 1904a. 'Objectivity' in social science. In *The methodology of the social sciences*, ed. E.A. Shils and H.A. Finch. New York: Free Press, 1959.

Weber, M. 1904b. *The protestant ethic and the spirit of capitalism*, , 1930. London: Allen & Unwin.

Weber, M. 1921. *Economy and society*, , 1968. New York: Bedminster Press.

Identification

Jean-Marie Dufour and Cheng Hsiao

Abstract

The problem of identification is defined in terms of the possibility of characterizing parameters of interest from observable data. This problem occurs in many fields, such as automatic control, biomedical engineering, psychology, systems science, the design of experiments, and econometrics. This article focuses on identification in econometric models, which typically involve random variables. Identification in general parametric statistical models is defined, and its meaning in a number of specific econometric models is considered: regression (collinearity), simultaneous equations, dynamic models, and nonlinear models. Identification in nonparametric models, weak identification, and the statistical implications of identification failure are also discussed.

Keywords

Bayes' th; Collinearity; Endogeneity and exogeneity; Identification; Instrumental variable; Linear models; Multivariate regression models; Nonparametric estimation; Nonparametric models; Probability; Random variables; Returns to schooling; Separability; Serial correlation; Simultaneous equations models; Treatment effect; Weak identification; Weak instruments

JEL Classifications

C3

In economic analysis, we often assume that there exists an underlying structure which has generated the observations of real-world data. However, statistical inference can relate only to characteristics of the distribution of the observed variables. Statistical models which are used to

explain the behaviour of observed data typically involve parameters, and statistical inference aims at making statements about these parameters. For that purpose, it is important that different values of a parameter of interest can be characterized in terms of the data distribution. Otherwise, the problem of drawing inferences about this parameter is plagued by a fundamental indeterminacy and can be viewed as 'ill-posed'.

To illustrate, consider X as being normally distributed with mean $E(X) = \mu_1 - \mu_2$. Then $\mu_1 - \mu_2$ can be estimated using observed X. But the parameters μ_1 and μ_2 are not uniquely estimable. In fact, one can think of an infinite number of pairs (μ_i, μ_j), $i, j = 1, 2, \ldots (i \neq j)$, such that $\mu_i - \mu_j = \mu_1 - \mu_2$. In order to determine μ_1 and μ_2 uniquely, we need additional prior information, such as $\mu_2 = 3\mu_1$ or some other assumption. Note, however, that inference about the variance of X remains feasible without extra assumptions.

More generally, *identification failures* –or situations that are close to it – complicate considerably the statistical analysis of models, so that tracking such failures and formulating restrictions to avoid them is an important problem of econometric modelling.

The problem of whether it is possible to draw inferences from the probability distribution of the observed variables to an underlying theoretical structure is the concern of econometric literature on identification. The first economists to raise this issue were Working (1925, 1927) and Wright (1915, 1928). The general formulations of the identification problems were made by Frisch (1934), Marschak (1942), Haavelmo (1944), Hurwicz (1950), Koopmans and Reiersøl (1950), Koopmans et al. (1950), Wald (1950), and many others. An extensive treatment of the theory of identification in simultaneous equation systems was provided by Fisher (1976). Surveys of the subject can be found in Hsiao (1983), Prakasa Rao (1992), Bekker and Wansbeek (2001), Manski (2003), and Matzkin (2007); see also Morgan (1990) and Stock and Trebbi (2003) on the early development of the subject.

In this article, we first define the notion of identification in general parametric models (Sections "Definition of Parametric Identification" and "General Results for Identification in Parametric Models") and discuss its meaning in a number of specific statistical models used in econometrics, such as regression models (collinearity), simultaneous equations, dynamic models, and nonlinear models (Section "Some Specific Parametric Models"). Identification in nonparametric models (Sections "Definition of Identification in Nonparametric Models" and "Examples of Nonparametric Identification"), weak identification (Section "Weak Instruments and Weak Identification"), and the statistical implications of identification failure (Section "Statistical Consequences of Identification Failure") are also considered.

Definition of Parametric Identification

It is generally assumed in econometrics that economic variables whose formation an economic theory is designed to explain have the characteristics of random variables. Let y be a set of such observations. A structure S is a complete specification of the probability distribution function of y. The set of all a priori possible structures, T, is called a model. In most applications, y is assumed to be generated by a parametric probability distribution function $F(y, \theta)$, where the probability distribution function F is assumed known, but the $q \times 1$ parameter vector θ is unknown. Hence, a structure is described by a parametric point θ, and a model is a set of points $A \subseteq \mathbb{R}^q$.

Definition 1 *Two structures, $S^0 = F(y, \theta^0)$ and $S^* = F(y, \theta^*)$ are said to be observationally equivalent if $F(y, \theta^0) = F(y, \theta^*)$ for ('almost') all possible y. A model is identifiable if A contains no two distinct structures which are observationally equivalent. A function of θ, g(θ), is identifiable if all observationally equivalent structures have the same value for g(θ).*

Sometimes a weaker concept of identifiability is useful.

Definition 2 *A structure with parameter value θ^0 is said to be locally identified if there exists an open neighborhood of θ^0, W, such that no other θ in W is observationally equivalent to θ^0.*

General Results for Identification in Parametric Models

Lack of identification reflects the fact that a random variable has the same distribution for some if not all values of the parameter. R.A. Fisher's information matrix provides a sensitivity measure of the distribution of a random variable due to small changes in the value of the parameter point (Rao 1962). It can therefore be shown that, subject to regularity conditions, θ^0 is locally identified if and only if the information matrix evaluated at θ^0 is nonsingular (Rothenberg 1971).

It is clear that unidentified parameters cannot be consistently estimated. There are also pathological cases where identified models fail to possess consistent estimators (for example, Gabrielson 1978). However, in most practical cases, we may treat identifiability and the existence of a consistent estimator as equivalent; for precise conditions, see Le Cam (1956) and Deistler and Seifert (1978).

Some Specific Parametric Models

The choice of model structure is one of the basic ingredients in the formulation of the identification problem. In this section we briefly discuss some identification conditions for different types of models in order to demonstrate the kind of prior restrictions required.

Linear Regression with Collinearity

One of the most common models where an identification problem does occur is the linear regression model:

$$\mathbf{y} = X\beta + \mathbf{u} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of dependent observable variables, X is an $n \times k$ fixed matrix of observable variables, β a $k \times 1$ unknown coefficient vector, and $\mathbf{u}$ is an $n \times 1$ vector of disturbances whose components are (say) independent and identically distributed according to a normal distribution $N(0, \sigma^2)$ with unknown positive variance σ^2.

In this model, the value of β must be determined from the expected value of $\mathbf{y}$: $E(\mathbf{y}) = X\beta$. If the latter equation has a solution for β (that is, if the model is correct), the solution is unique if and only the regressor matrix X has rank k. If X has rank zero (which entails $X = 0$), all values of β are equivalent *(β is completely unidentifiable).* If $1 \leq \text{rank}(X) < k$, then not all the components can be determined, but some linear combinations of the components of β (say $c'\beta$) can be determined (that is, they are *identifiable*). A necessary and sufficient condition for $c'\beta$ to be estimable (identifiable) is that $c = (X'X)d$ for some vector d. Linear combinations that do not satisfy this condition are not identifiable. The typical way out of such collinearity problems consists in imposing restrictions on β (identifying restrictions) which set the values of the unidentifiable linear combinations (or components) of β.

Correspondingly, when X does not have full rank, the equation $(X'X)\hat{\beta} = X'\mathbf{y}$, which defines the least squares estimator $\hat{\beta}$, does not have a unique solution. But all solutions of the least squares problem can be determined by considering $\hat{\beta} = (X'X)^{-}X'\mathbf{y}$ where $(X'X)^{-}$ is any generalized inverse of $(X'X)$. Different generalized inverses then correspond to different identifying restrictions on β. For further discussion, see Rao (1973, ch. 4).

Linear Simultaneous Equations Models

Consider a theory which predicts a relationship among the variables as

$$B\mathbf{y}_t + \Gamma\mathbf{x}_t = \mathbf{u}_t, \quad t = 1, \ldots, n, \tag{2}$$

where $\mathbf{y}_t$ and $\mathbf{u}_t$, are $G \times 1$ vectors of observed and unobserved random variables, respectively, $\mathbf{x}_t$ is a $K \times 1$ vector of observed non-stochastic variables, B and Γ are $G \times G$ and $G \times K$ matrices of coefficients, with B nonsingular. We assume that the $\mathbf{u}_t$ are independently normally distributed with mean 0 and variance-covariance matrix $\sum$. Equations (2) are called structural equations. Solving for the endogenous variables, $\mathbf{y}$, as a function of the exogenous variables, $\mathbf{x}$, and the disturbance u, we obtain:

$$\mathbf{y}_t = -B^{-1}\Gamma\mathbf{x}_t + B^{-1}\mathbf{u}_t = \Pi\mathbf{x}_t + \mathbf{v}_t, \quad (3)$$

Where $\Pi = -B^{-1}\Gamma, E\mathbf{v}_t = 0, E\mathbf{v}_t\mathbf{v}_t' = V = B^{-1}\sum\left(B^{-1}\right)'$. Equations (3) are called the *reduced form* equations derived from (2) and give the conditional likelihood of $\mathbf{y}_t$ for given $\mathbf{x}_t$ that summaries the information provided by the observed ($\mathbf{y}_t$, $\mathbf{x}_t$,). The variables in $\mathbf{x}_t$ are often also called 'instruments'.

From (3), we see that the simultaneous equations model can be viewed as a special case of a multivariate regression model (MLR), such that the regression coefficient matrix Π satisfies the equation:

$$B\Pi = -\Gamma. \quad (4)$$

Provided the matrix $X = [x_1, \ldots, x_n]'$ has full rank K (no collinearity), the regression coefficient matrix Π is uniquely determined by the distribution of $Y = [y_1, \ldots, y_n]'$ (it is *identifiable)*. The problem is then whether B and Γ can be uniquely derived from Eq. (4). Premultiplying (2) by a $G \times G$ nonsingular matrix D, we get a second structural equation:

$$B^*\mathbf{y}_t + \Gamma^*\mathbf{x}_t = \mathbf{u}_t^*, \quad (5)$$

where $B^* = DB$, $\Gamma^* = D\Gamma$, and $\mathbf{u}_t^* = D\mathbf{u}$. It is readily seen that the reduced form of (5) is also (3). So Eq. (4) cannot be uniquely solved for B and Γ, given Π. Therefore, the two structures are observationally equivalent and the model is *non-identifiable.*

To make the model identifiable, additional prior restrictions have to be imposed on the matrices B, Γ and/or $\sum$. Consider the problem of estimating the parameters of the first equation in (2), out of a system of G equations. If the parameters cannot be estimated, the first equation is called *unidentified* or *underidentified*. If given the prior information, there is a unique way of estimating the unknown parameters, the equation is called *just identified.* If the prior information allows the parameters to be estimated in two or more linearly independent ways, it is called *over-identified.* A necessary condition for the first equation to be identified is that the number of restrictions on this equation be no less than $G - 1$ (order condition). A necessary and sufficient condition is that a specified submatrix of B, Γ and $\sum$ be of rank $G - 1$ (rank condition) (see Fisher 1976; Hausman and Taylor 1983). For instance, suppose the restrictions on the first equation are in the form that certain variables do not appear. Then this rank condition says that the first equation is identified if and only if the submatrix obtained by taking the columns of B and Γ with prescribed zeros in the first row is of rank $G - 1$ (Koopmans and Reiersøl 1950).

Dynamic Models

When both lagged endogenous variables and serial correlation in the disturbance term appear, we need to impose additional conditions to identify a model. For instance, consider the following two equation system (Koopmans et al. 1950):

$$y_{1t} + \beta_{11}y_{1,t-1} + \beta_{12}y_{2,t-1} = u_{1t}, \quad \beta_{12}y_{1t} + y_{2t} = u_{2t}. \quad (6)$$

If (u_{1t}, u_{2t}) are serially uncorrelated, (6) is identified. If serial correlation in (u_{1t}, u_{2t}) is allowed, then

$$y_{1t} + \beta^*{}_{11}y_{1,t-1} + \beta^*{}_{12}y_{2,t-1} = u^*{}_{1t}, \quad \beta_{12}y_{1t} + y_{2t} = u_{2t}, \quad (7)$$

is observationally equivalent to (6), where $\beta_{11}^* = \beta_{11} + d\beta_{21}$, $\beta_{12}^* = \beta_{12} + d$, and $u_{1t}^* = u_{1t} + du_{2t}$.

Hannan (1971) derives generalized rank conditions for the identification of this type of model by first assuming that the maximum orders of lagged endogenous and exogenous variables are known, then imposing restrictions to eliminate redundancy in the specification and to exclude transformations of the equations that involve shifts in time. Hatanaka (1975), on the other hand, assumes that the prior information takes only the form of excluding certain variables from an equation, and derives a rank condition which allows common roots to appear in each equation.

I

Nonlinear Models

For linear models, we have either global identification or else an infinite number of observationally equivalent structures. For models that are linear in parameters, but nonlinear in variables, there is a broad class of models whose members can commonly achieve identification (Brown 1983; McManus 1992). For models linear in the variables but nonlinear in the parameters, the state of the mathematical art is such that we only talk about local properties. That is, we cannot tell the true structure from any other substitute; however, we may be able to distinguish it from other structures which are close to it. A sufficient condition for local identification is that the Jacobian matrix formed by taking the first partial derivatives of

$$\omega_i = \Psi_i(\theta), i = 1, \ldots, n, 0 = \phi_j(\theta), \; j = 1, \ldots, R, \tag{8}$$

with respect to θ be of full column rank, where the ω_i are n population moments of y and the ϕ_j are the R a priori restrictions on θ (Fisher 1976).

When the Jacobian matrix of (8) has less than full column rank, the model may still be locally identifiable via conditions implied by the higher-order derivatives. However, the estimator of a model suffering from first-order lack of identification will in finite samples behave in a way which is difficult to distinguish from the behaviour of an unidentified model (Sargan 1983).

Bayesian Analysis

In Bayesian analysis all quantities, including the parameters, are random variables. Thus, a model is said to be identified in probability if the posterior distribution for θ is proper. When the prior distribution for θ is proper, so is the posterior, regardless of the likelihood function of y. In this sense unidentifiability causes no real difficulty in the Bayesian approach. However, basic to the Bayesian argument is that all probability statements are conditional, that is, they consist essentially in revising the probability of a fixed event in the light of various conditioning events, the revision being accomplished by Bayes' theorem. Therefore, in order for an experiment to be informative with regard to unknown parameters (that is, for the posterior to be different from the prior), the parameter must be identified or estimable in the classical sense and identification remains as a property of the likelihood function (Kadane 1975).

Drèze (1975) has commented that exact restrictions are unlikely to hold with probability 1 and has suggested using probabilistic prior information. In order to incorporate a stochastic prior, he has derived necessary rank conditions for the identification of a linear simultaneous equation model.

Definition of Identification in Nonparametric Models

When the restrictions of an economic model specify all functions and distributions up to the value of a finite dimensional vector, the model is said to be parametric. When some functions or distributions are left parametrically unspecified, the model is said to be semiparametric. The model is nonparametric if none of the functions and distributions are specified parametrically. The previous discussion is based on parametric specification. We now turn to the issue of whether economic restrictions such as concavity, continuity and monotonicity of functions, equilibrium conditions, the implications of optimization, and so on, may be used to guarantee the identification of some nonparametric models and the consistency of some nonparametric estimators (see Matzkin 1994).

Formally, an econometric model is specified by a vector of observable dependent and independent variables, a vector of unobservable variables, and a set of known functional relationships among the variables. When such functional relationships are unspecified, the nonparametric identification studies what functions or features of function can be recovered from the joint distribution of the observable variables.

The set of restrictions on the unknown functions and distributions in an econometric model

defines the set of functions and distributions to which these belong. Let the model T denote the set of all a priori possible unknown functions and distributions. Let m denote a vector of the unknown functions and distributions in T and $P(m)$ denote the joint distribution of the observable variables under m. Then the identification of m can be defined as follows.

Definition 3 *The vector of functions m is identified in T if for any other vector, m* ∈ T such that* $m \neq m^*$, $P(m) \neq P(m^*)$.

Let $C(m)$ denote some feature of m, such as the sign of some coordinate of m.

Definition 4 *The feature C(m) of m is identified if* $C(m) = C(m^*)$ *for all m,* $m^* \in T$ *such that* $P(m) = P(m^*)$.

Examples of Nonparametric Identification

Contrary to the parametric model, there is no general result for nonparametric identification. We shall therefore give some examples of how restrictions can be used to identify nonparametric functions.

Generalized Regression Models

Economists often consider a model of the form

$$y = g(\mathbf{x}) + u. \tag{9}$$

When $E(u|\mathbf{x}) = 0$ and $g(\cdot)$ is a continuous function $g : x \to \mathbb{R}$, then $g(\cdot)$ can be recovered from the joint distribution of (y, **x**) because $E(y|\mathbf{x}) = g(\mathbf{x})$.

In some cases, the object of interest is not a conditional mean function g(.), but some 'deeper' function, such as a utility function generating the distribution of demand for commodities by a consumer. For example, **x** in (9) can be a price vector for K commodities and the income of a consumer. Mas-Colell (1977) has shown that we can recover the underlying utility function from the distribution of demand if we restrict $g(\cdot)$ to be monotone increasing, continuous, concave and strictly quasiconcave functions.

Simultaneous Equations Models

Suppose (**y**, **x**) satisfies the structural equations

$$\mathbf{r}(\mathbf{x}, \mathbf{y}) = \mathbf{u}, \tag{10}$$

where **y** and **u** denote $G \times 1$ vectors of observable endogenous and unobservable variables, respectively, **x** is a $K \times 1$ vector of observable exogenous variables, **r** denotes the G unknown functions, and let $p(\mathbf{r})$ and $p(\mathbf{r}^*)$ represent the joint distributions of the observables under **r** and **r*** respectively. Assume also that: (i) $\forall(\mathbf{x}, \mathbf{y})$, $\partial\mathbf{r}/\partial\mathbf{y}$ has full rank, (ii) there exists a function $\pi(\cdot)$ such that $y = \pi(\mathbf{x}, \mathbf{u})$ (for conditions ensuring this, see Benkard and Berry 2006), and (iii) **u** is distributed independently of **x**. Then a necessary and sufficient condition guaranteeing that $p(\mathbf{r}^*) = p(\mathbf{r})$ is that

$$\operatorname{rank} \begin{pmatrix} \dfrac{\partial \mathbf{r}_i^*}{\partial(\mathbf{x}, \mathbf{y})} \\ \dfrac{\partial \mathbf{r}}{\partial(\mathbf{x}, \mathbf{y})} \end{pmatrix} < G + 1, \tag{11}$$

for all (**x**, **y**) and $i = 1,\ldots,G$, and all, where $\mathbf{r}_i^*$ denotes the i-th coordinate function of $\mathbf{r}^* \in T$ (see Roehrig 1988; Matzkin 2007).

Latent Variable Models and the Measurement of Treatment Effects

For each person i, let (y_{0i}^*, y_{1i}^*) denote the potential outcomes in the untreated and treated states, respectively. Then the treatment effect for individual i is

$$\Delta_i = y_{1i}^* - y_{0i}^*$$

and the average treatment effect (ATE) is defined as

$$E(\Delta_i) = E\left(y_{1i}^* - y_{0i}^*\right); \tag{12}$$

see Heckman and Vytlacil (2001).

Let the treatment status be denoted by the dummy variable d_i where $d_i = 1$ denotes the

receipt of treatment and $d_i = 0$ denotes nonreceipt. The observed data are often in the form

$$y_i = d_i y_{1i}^* + (1 - d_i) y_{0i}^*. \tag{13}$$

Suppose $y_{1i}^* = \mu_1(\mathbf{x}_i, u_{1i})$, $y_{0i}^* = \mu_0(\mathbf{x}_i, u_{0i})$ and $d_i^* = \mu_D(\mathbf{z}_i) - u_{di}$, where $d_i = 1$ if $d_i^* \geq 0$ and 0 otherwise, $\mathbf{x}_i$, and $\mathbf{z}_i$, are vectors of observable exogenous variables and (u_{1i}, u_{0i}, u_{di}) are unobserved random variables. The average treatment effect and the complete structural econometric model can be identified with parametric specifications of $(\mu_1(\cdot), \mu_0(\cdot), \mu_D(\cdot))$ and the joint distributions of (u_{1i}, u_{0i}, u_{di}) even though we do not simultaneously observe y_{1i}^* and y_{0i}^*. In the case that neither $(\mu_1(\cdot), \mu_0(\cdot), \mu_D(\cdot))$ nor the joint distribution of (u_1, u_0, u_d) are specified, certain treatment effects may still be nonparametrically identified under weaker assumptions. For instance, under the assumption that d_i is orthogonal to $\left(y_{1i}^*, y_{0i}^*\right)$ conditional on a set of confounders (x, z) (conditional independence or ignorable selection), the ATE is identifiable and estimable by comparing the difference of the average outcomes from the treatment group and from the untreated (control) group (Heckman and Robb 1985; Rosenbaum and Rubin 1985). If the focus is on the average treatment effect for someone who would not participate if $p(\mathbf{z}) \leq p(\mathbf{z}_0)$ and would participate if $p(\mathbf{z}) > p(\mathbf{z}_0)$ (the local average treatment effect (LATE)), where $p(\mathbf{z}) = \text{Prob}(d = 1|\mathbf{z})$ (propensity score), Imbens and Angrist (1994) show that under the assumptions of separability of the effects of observable factors and unobservable factors and independence between observed factors and unobserved factors, they can be estimated by the sample analogue of

$$\Delta^{LATE}(\mathbf{x}, p(\mathbf{z}), p(\mathbf{z}_0)) \equiv \frac{E(y|\mathbf{x}, p(\mathbf{z})) - E(y|\mathbf{x}, p(\mathbf{z}_0))}{p(\mathbf{z}) - p(\mathbf{z}_0)} \tag{14}$$

where, without loss of generality, we assume $p(\mathbf{z}) > p(\mathbf{z}_0)$. The limit of LATE provides the local instrumental variable (LIV) estimand (Heckman and Vytlacil 1999):

$$\Delta^{LIV}(\mathbf{x}, p(\mathbf{z})) \equiv \frac{\partial E(y|\mathbf{x}, p(\mathbf{z}))}{\partial p(\mathbf{z})}. \tag{15}$$

Heckman and Vytlacil (2001) give conditions that suitably weighted versions of LIV identify the ATE.

Weak Instruments and Weak Identification

The most common way of trying to achieve identification consists in imposing exclusion restrictions on the variables of a structural equation. In model (2), suppose that $\mathbf{y}_t$ and $\mathbf{x}_t$ are partitioned as $\mathbf{y}_t = \left(y_{1t}, \mathbf{y}_{2t}', \mathbf{y}_{3t}'\right)'$ and $\mathbf{x}_t = \left(\mathbf{x}_{1t}', \mathbf{x}_{2t}'\right)'$ where y_{1t} is a scalar, $\mathbf{y}_{it}$ has dimension $G_i (i = 2, 3)$ and $\boldsymbol{x}_{it}$ has dimension $K_i (i = 1, 2)$. If $\mathbf{y}_{3t}$ and $\mathbf{x}_{2t}$ are excluded from the first equation and the coefficient of y_{1t} is normalized to one, this yields an equation of the form:

$$y_{1t} + \mathbf{y}_{2t}'\beta_1 = \mathbf{x}_{1t}'\gamma_1 + u_{1t}, \quad t = 1, \ldots, n. \tag{16}$$

Let us also rewrite the reduced equation for $\mathbf{y}_{2t}$ in terms of $\mathbf{x}_{1t}$ and $\mathbf{x}_{2t}$:

$$\mathbf{y}_{2t} = \Pi_{21}\mathbf{x}_{1t} + \Pi_{22}\mathbf{x}_{2t} + \mathbf{v}_{2t}. \tag{17}$$

Then, substituting (17) into (16), we see that the reduced form for y_{1t} is:

$$y_{1t} = \Pi_{11}\mathbf{x}_{1t} + \Pi_{12}\mathbf{x}_{2t} + \mathbf{v}_{1t}, \tag{18}$$

where $\mathbf{v}_{1t} = u_{1t} + \mathbf{v}_{2t}'\beta_1$, $\Pi_{11} = \Upsilon_1' + \beta_1'\Pi_{21}$ and

$$\Pi_{12}' = \Pi_{22}'\beta_1. \tag{19}$$

Since γ_1 is free, Π_{11} is not restricted, but Eq. (19) determines the identifiability of β_1, hence also of γ_1. Provided Eq. (19) has a solution (that is, if Eq. (16) is consistent with the data), the solution is unique if and only if the rank of the $G_2 \times K_2$ matrix Π_{22} is equal to G_2, the dimension of β_1:

$$\text{rank}(\Pi_{22}) = G_2. \tag{20}$$

If $\text{rank}(\Pi_{22}) < G_2$, the vector β_1 is not identifiable. However, it is completely unidentifiable only if $\text{rank}(\Pi_{22}) = 0$, or equivalently if $\Pi_{22} = 0$. If $1 < \text{rank}(\Pi_{22}) < G_2$, some linear combinations $c'\beta_1$ are identifiable, but not all of them. Failure of the identification condition means that the regressors (or the 'instruments') $\mathbf{x}_{2t}$ do not move enough to separate the effects of the different variables in $\mathbf{y}_{2t}$. Condition (20) underscores two important things: first, exclusion and normalization restrictions – which are easy to check – are not sufficient to ensure identification; second, identification depends on the way the exogenous variables $\mathbf{x}_{2t}$ *excluded* from the structural equation of interest (16) are related to endogenous variables $\mathbf{y}_{2t}$ included in the equation. The latter feature is determined by the matrix Π_{22} whose rows should be linearly independent. Since Π_{22} is not observable, this may be difficult to determine in practice.

A situation that can lead to identification difficulties is the one where the identification condition (20) indeed holds, but, in some sense, Π_{22} is 'close' not to have sufficient rank. In such situations, we say that we have *weak instruments.* In view of the fact that the distributions of most statistics move continuously as functions of Π_{22}, the practical consequences of being close to identification failure are essentially the same. Assessing the closeness to non-identification may be done in various ways, for example by considering the eigenvalues of the matrices which measure the 'size' of Π_{22}, such as $\Pi_{22}\Pi'_{22}$, $\Pi_{22}X'_2M(X_1)X_2\Pi'_{22}$, or a *concentration matrix* $\sum_{22}^{-1/2}\Pi_{22}X'_2M(X_1)X_2\Pi'_{22}\sum_{22}^{-1/2}$ where $X_1 = [\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n}]'$, $X_2 = [\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n}]'$, $\sum_{22}$ is the covariance matrix of v_{2t}, $\sum_{22}^{-1/2}$ is its square root, and $M(X_1) = I_n - X_1\left(X'_1X_1\right)^{-1}X'_1$. More generally, any situation where a parameter may be difficult to determine because we are close to a case where a parameter ceases to be identifiable may be called *weak identification.* Weak identification was highlighted as a problem of practical interest by Nelson and Startz (1990), Bound et al. (1995), Dufour (1997), and Staiger and Stock (1997); for reviews, see Stock et al. (2002) and Dufour (2003).

Statistical Consequences of Identification Failure

Identification failure has several detrimental consequences for statistical analysis:

1. Parameter estimates, tests and confidence sets computed for unidentified parameters have no clear input; this situation may be especially misleading if the statistical instruments used do not reveal the presence of the problem.
2. Consistent estimation is not possible unless additional information is supplied.
3. Many standard distributional results used for inference on such models are not anymore valid, even with a large sample size (see Phillips 1983, 1989; Rothenberg 1984).
4. Numerical problems also easily appear, due for example to the need to invert (quasi) singular matrices.

Weak identification problems lead to similar difficulties, but may be more treacherous in the sense that standard asymptotic distributional may remain valid, but they constitute very bad approximations to what happens in finite samples:

1. Standard consistent estimators of structural parameters can be heavily biased and follow distributions whose form is far from the limiting Gaussian distribution, such as bimodal distributions, even with fairly large samples (Nelson and Startz 1990; Hillier 1990; Buse 1992).
2. Standard tests and confidence sets, such as Wald-type procedures based on estimated standard errors, become highly unreliable or completely invalid (Dufour 1997).

A striking illustration of these problems appears in the reconsideration by Bound et al.

(1995) of a study on returns to education by Angrist and Krueger (1991). Using 329,000 observations, these authors found that replacing the instruments used by Angrist and Krueger (1991) with randomly generated (totally irrelevant) instruments produced very similar point estimates and standard errors.

This result indicates that the original instruments were weak. Recent work in this area is reviewed in Stock et al. (2002) and Dufour (2003).

Concluding Remarks

The study of identifiability is undertaken in order to explore the limitations of statistical inference (when working with economic data) or to specify what sort of a priori information is needed to make a model estimable. It is a fundamental problem concomitant with the existence of a structure. Logically it precedes all problems of estimation or of testing hypotheses.

An important point that arises in the study of identification is that without a priori restrictions imposed by economic theory it would be almost impossible to estimate economic relationships. In fact, Liu (1960) and Sims (1980) have argued that economic relations are not identifiable because the world is so interdependent as to have almost all variables appearing in every equation, thus violating the necessary condition for identification. However, almost all the models we discuss in econometrics are only approximate. We use convenient formulations which behave in a general way that corresponds to our economic theories and intuitions, and which cannot be rejected by the available data. In this sense, identification is a property of the model but not necessarily of the real world. It is also important to be careful about situations where identification almost does not hold (weak identification), since these are in practice as damaging for statistical analysis as identification failure itself.

The problem of identification arises in a number of different fields such as automatic control, biomedical engineering, psychology, systems science, and so on, where the underlying physical structure may be deterministic (for example, see Aström and Eykhoff 1971). It is also aptly linked to the design of experiments (for example, Kempthorne 1947; Bailey et al. 1977). Here, we restrict our discussion to economic applications of statistical identifiability involving random variables.

See Also

- ▶ Econometrics
- ▶ Endogeneity and Exogeneity
- ▶ Simultaneous Equations Models
- ▶ Treatment Effect

Bibliography

Angrist, J.D., and A.B.. Krueger. 1991. Does compulsory school attendance affect schooling and earning? *Quarterly Journal of Economics* 106: 979–1014.

Aström, K.J., and P. Eykhoff. 1971. System identification – A survey. *Automatica* 7: 123–162.

Bailey, R.A., F.H.L. Gilchrist, and H.D. Patterson. 1977. Identification of effects and confounding patterns in factorial designs. *Biometrika* 64: 347–354.

Bekker, P., and T. Wansbeek. 2001. Identification in parametric models. In *Companion to theoretical econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Benkard, C.L., and S. Berry. 2006. On the nonparametric identification of nonlinear simultaneous equations models: Comment on Brown (1983) and Roehrig (1988). *Econometrica* 74: 1429–1440.

Bound, J., D.A. Jaeger, and R.M. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90: 443–450.

Brown, B.W. 1983. The identification problem in systems nonlinear in the variables. *Econometrica* 51: 175–196.

Buse, A. 1992. The bias of instrumental variables estimators. *Econometrica* 60: 173–180.

Deistler, M., and H.-G. Seifert. 1978. Identifiability and consistent estimability in econometric models. *Econometrica* 46: 969–980.

Drèze, J. 1975. Bayesian theory of identification in simultaneous equations models. In *Studies in Bayesian econometrics and statistics*, ed. S.E. Fienberg and A. Zellner. Amsterdam: North-Holland.

Dufour, J.-M. 1997. Some impossibility theorems in econometrics, with applications to structural and dynamic models. *Econometrica* 65: 1365–1389.

Dufour, J.-M. 2003. Identification, weak instruments and statistical inference in econometrics. *Canadian Journal of Economics* 36: 767–808.

Fienberg, S.E., and A. Zellner, eds. 1975. *Studies in Bayesian econometrics and statistics*. Amsterdam: North-Holland.

Fisher, F.M. 1976. *The identification problem in econometrics*. Huntington: Krieger.

Frisch, R. 1934. *Statistical confluence analysis by means of complete regression systems*. Oslo: Universitetes Okonomiske Institutt.

Gabrielson, A. 1978. Consistency and identifiability. *Journal of Econometrics* 8: 261–263.

Griliches, Z., and M.D. Intriligator, eds. 1983. *Handbook of econometrics*, vol. 1. Amsterdam: North-Holland.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12 (Supp): 1–115.

Hannan, E.J. 1971. The identification problem for multiple equation systems with moving average errors. *Econometrica* 39: 751–766.

Hatanaka, M. 1975. On the global identification of the dynamic simultaneous equations model with stationary disturbances. *International Economic Review* 16: 545–554.

Hausman, J.A., and W.E. Taylor. 1983. Identification, estimation and testing in simultaneous equations models with disturbance covariance restriction. *Econometrica* 51: 1527–1549.

Heckman, J., and R. Robb. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, ed. J. Heckman and B. Singer. New York: Cambridge University Press.

Heckman, J.J., and E. Vytlacil. 1999. Local instrumental variables and latent variables models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences* 96: 4730–4734.

Heckman, J.J., and E. Vytlacil. 2001. Local instrumental variables. In *Nonlinear statistical modeling proceedings of the thirteenth international symposium in economic theory and econometrics: Essays in honor of Takeshi Amemiya*, ed. C. Hsiao, K. Morimune, and J.L. Powell. Cambridge: Cambridge University Press.

Hillier, G.H. 1990. On the normalization of structural equations: Properties of direction estimators. *Econometrica* 58: 1181–1194.

Hsiao, C. 1983. Identification. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 1. Amsterdam: North-Holland.

Hurwicz, L. 1950. Generalization of the concept of identification. In *Statistical inference in dynamic economic models*, ed. T.C. Koopmans. New York: Wiley.

Imbens, G., and J. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62: 467–476.

Kadane, J.B. 1975. The role of identification in Bayesian theory. In *Studies in Bayesian econometrics and statistics*, ed. S.E. Fienberg and A. Zellner. Amsterdam: North-Holland.

Kempthorne, O. 1947. A simple approach to confounding and factorial replication in factorial experiments. *Biometrika* 34: 255–272.

Koopmans, T.C. 1950. *Statistical inference in dynamic economic models*. New York: Wiley.

Koopmans, T.C., and O. Reiersøl. 1950. The identification of structural characteristics. *Annals of Mathematical Statistics* 21: 165–181.

Koopmans, T.C., H. Rubin, and R.B. Leipnik. 1950. Measuring the equation systems of dynamic economics. In *Statistical inference in dynamic economic models*, ed. T.C. Koopmans. New York: Wiley.

Le Cam, L. 1956. On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the third Berkeley Symposium on mathematical statistics and probability*. Berkeley: University of California Press.

Liu, T.C. 1960. Underidentification, structural estimation, and forecasting. *Econometrica* 28: 855–865.

Manski, C. 2003. *Partial identification of probability distributions*. New York: Springer.

Marschak, J. 1942. Economic interdependence and statistical analysis. In *Studies in mathematical economics and econometrics*, ed. O. Lange, F. McIntyre, and T.O. Yntema. Chicago: University of Chicago Press.

Mas-Collel, A. 1977. On the recoverability of consumers preferences from market demand behavior. *Econometrica* 45: 1409–1430.

Matzkin, R. 1994. Restrictions of economic theory in nonparametric methods. In *Handbook of econometrics*, ed. R.F. Engle and D.L. McFadden, vol. 4. Amsterdam: North-Holland.

Matzkin, R. 2007. Nonparametric identification. In *Handbook of econometrics*, ed. J. Heckman and E. Leamer, vol. 6. Amsterdam: North-Holland.

McManus, D.A. 1992. How common is identification in parametric models? *Journal of Econometrics* 53: 5–23.

Morgan, M.S. 1990. *The history of econometric ideas*. Cambridge: Cambridge University Press.

Nelson, C.R., and R. Startz. 1990. The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one. *Journal of Business* 63: 125–140.

Phillips, P.C.B. 1983. Exact small sample theory in the simultaneous equations model. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 1. Amsterdam: North-Holland.

Phillips, P.C.B. 1989. Partially identified econometric models. *Econometric Theory* 5: 181–240.

Prakasa Rao, B.L.S. 1992. *Identifiability in stochastic models: Characterization of probability distributions*. New York: Academic Press.

Rao, C.R. 1962. Problems of selection with restriction. *Journal of the Royal Statistical Society, Series B* 24: 401–405.

Rao, C.R. 1973. *Linear statistical inference and its applications,* 2nd ed. New York: Wiley.

Roehrig, C.S. 1988. Conditions for identification in nonparametric and parametric models. *Econometrica* 56: 433–477.

Rosenbaum, P., and D. Rubin. 1985. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79: 516–524.

Rothenberg, T.J. 1971. Identification in parametric models. *Econometrica* 39: 577–591.

Rothenberg, T.J. 1984. Approximating the distributions of econometric estimators and test statistics. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.

Sargan, J.D. 1983. Identification and lack of identification. *Econometrica* 51: 1605–1633.
Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.
Staiger, D., and J.H. Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65: 557–586.
Stock, J.H., and F. Trebbi. 2003. Who invented IV regression? *Journal of Economic Perspectives* 17 (3): 177–194.
Stock, J.H., J.H. Wright, and M. Yogo. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20: 518–529.
Wald, A. 1950. Note on the identification of economic relations. In *Statistical inference in dynamic economic models*, ed. T.C. Koopmans. New York: Wiley.
Working, H. 1925. The statistical determination of demand curves. *Quarterly Journal of Economics* 39: 503–543.
Working, E.J. 1927. What do statistical demand curves show? *Quarterly Journal of Economics* 41: 212–235.
Wright, P.G. 1915. Moore's economic cycles. *Quarterly Journal of Economics* 29: 631–641.
Wright, P.G. 1928. *The tariff on animal and vegetable oils*. New York: Macmillan.

Identity

George Akerlof and Rachel Kranton

Keywords
Gender roles; Identity; Signalling; Social norms; Well-being

JEL Classifications
D11

A person's *identity* is broadly defined as a person's self-image or sense of self. The concept of identity has wide use in most social sciences outside economics, especially sociology, anthropology and psychology. Many social scientists hold that preserving or enhancing identity is a prime motivation for individual and group behaviour. At the time of this writing, economists are beginning to explore the implications of identity for economic outcomes.

To do so, researchers primarily include identity as an aspect of utility. In this view, a person's actions and consumption of goods and services not only affect their material well-being, but also their psychological well-being. Researchers then ask how the inclusion of identity in utility can affect economic outcomes, such as charitable contributions (Bénabou and Tirole 2006), information acquisition (Kőszegi 2006), schooling rates (Akerlof and Kranton 2002), and the design of workplace incentives (Akerlof and Kranton 2005).

We can divide the economic research on identity into two strands. The first considers an individual's self-image, as in Bénabou and Tirole (2005) and Kőszegi (2006). The second considers an individual's self-image as it relates to societal norms and ideals (Akerlof and Kranton 2000, 2002, 2005).

The first strand of research explores the simple proposition that people like to feel good about themselves. There are then trade-offs between standard economic costs and benefits, and the costs and benefits for one's own self-image. Kőszegi (2006) uses such a utility function to explain why people may not undertake profitable investment projects, as the downside payoffs also reduce a person's sense of his own abilities. Bénabou and Tirole (2005) use identity to explain why monetary compensation can reduce the levels of pro-social activities (such as volunteer work and blood donations), as found in several studies and experimental work. They posit a utility function where an individual's action yields a monetary payoff and an 'intrinsic' payoff. Individuals can have different valuations/preferences for the monetary payoff and the intrinsic payoff. Individuals like to think of themselves as placing, and like to think others think they place, high values on intrinsic payoff. That is, they want to think of themselves as enjoying the pro-social action for its own sake. But preferences are not observable, perhaps even to oneself. As individuals choose different actions, they and others make inferences about preferences. Hence, actions serve as 'self-signal' and a signal to others. The main results concern the trade-off between monetary payoffs and the signalling value of an action. When the

monetary compensation for an action increases, the signal conveys less information about a person's underlying value for intrinsic payoffs. Hence, introducing monetary rewards can lower the levels of pro-social activity.

The second strand of research considers identity and norms. Sen (1985) and Elster (1989) were among the earliest proponents of the importance for economics of utility-based norms. Akerlof and Kranton (2000, 2002, 2005) relate a person's self-image to societal norms and ideals for different people in society. Whether or not a person feels good about herself depends on how that person *should* act, according to her place in society. Thus, to take the most obvious example, men are supposed to act differently from women, and identity utility will depend on the match between a person's actions and these gender norms. This notion of identity reflects a large body of research on 'social identity' in psychology, reviewed in Haslam (2001). Philosophy has also been another important influence on the connection between identity and norms, especially for Elster (1989) and Sen (1985).

Akerlof and Kranton (2000) posits the following utility function for an individual j:

$$U_j(a_j, a_{-j}, I_j)$$

where a_j are j's actions, a_{-j} are others' actions, and I_j is j's 'identity utility' which is itself a function:

$$I_j(a_j, a_{-j}; c_j, \varepsilon_j, N)$$

where c_j denotes j's *social category*, N denotes the norms of behaviour and ideal attributes for different social categories, and ε_j denotes j's own attributes. The inclusion of others' actions allows for identity externalities. In the simplest case, an individual j chooses actions a_j to maximize utility U, taking as given c_j, ε_j, and N and the actions of others. In some applications, individuals may also choose the category assignment c_j, as social categories may be more or less ascriptive. Individual actions may also affect the norms, N, the set of social categories, C, as well as the status of different categories reflected in $I_j(\cdot)$. With respect to gender, for example, the women's movement strived to reduce status differences between men and women and change prescribed behaviour. Gender categories themselves have become varied and complex over time. There may be no universal agreement about social categories and prescriptions. Indeed, they are the subject of much debate and controversy and the source of new externalities.

This utility function highlights a different motivation for behaviour from a standard model, and shows how social identity can affect economic outcomes. For example, in the workplace different workers may feel more or less part of an organization (*insiders* versus *outsiders*), and work incentives will depend on norms for these different categories of workers. This utility function has implications for supervisory and management policy, as in Akerlof and Kranton (2005). A firm could choose a strict supervisory policy where a supervisor reports to upper management on workers' behaviour. This policy yields greater information, but can lead to workers adopting an *outsider* identity, with lower work norms. A looser supervisory policy yields less information to management, but workers develop a work group identity with possibly higher work norms. We use our utility function to explore the implications of identity in other realms, including race and poverty (Akerlof and Kranton 2000), gender in the labour market (Akerlof and Kranton 2000), and schools, student identity and education (Akerlof and Kranton 2002).

See Also

- ▶ Culture and Economics
- ▶ Social Norms

Bibliography

Akerlof, G., and R. Kranton. 2000. Economics and identity. *Quarterly Journal of Economics* 115: 715–733.

Akerlof, G., and R. Kranton. 2002. Identity and schooling: Some lessons for the economics of education. *Journal of Economic Literature* 40: 1167–1201.

I

Akerlof, G., and R. Kranton. 2005. Identity and the economics of organizations. *Journal of Economic Perspectives* 19(1): 9–32.
Bénabou, R., and J. Tirole. 2006. Incentives and prosocial behavior. *American Economic Review* 96(5): 1652–1678.
Elster, Jon. 1989. Social norms and economic theory. *Journal of Economic Perspectives* 3(4): 99–117.
Haslam, S. 2001. *Psychology in organizations: The social identity approach*. London/Thousand Oaks: Sage.
Kőszegi, B. 2006. Ego utility, overconfidence, and task choice. *Journal of the European Economic Association* 4: 673–707.
Sen, A. 1985. Goals, commitment, and identity. *Journal of Law, Economics, and Organization* 1: 341–355.

Ideology

Kurt Klappholz

Now and then one comes across the claim that, unlike, for example, physics, 'economics is thoroughly permeated by ideology . . .' (Ward 1979, p. viii). The exact import of this claim regarding the epistemological status of economics is not clear, since the noun 'ideology' is employed in a variety of senses. However, it should be stressed at once that, despite occasional criticisms (e.g. McCloskey 1983, p. 334), most economists long ago accepted Hume's insistence that policy proposals cannot be deduced from descriptive statements alone (Klappholz 1964) and have therefore stressed the distinction between positive and normative economics. The claim discussed in this essay appears to be directed at both the positive, as well as the normative, parts of economics, but we shall be concerned mainly with its import for positive economics. In section I we interpret the claim that economics is ideological as the view that economic theories can be explained by the social position and attitudes of those who put them forward, that is, by the Sociology of Knowledge (discussed critically in Popper 1957, chs 23 and 24). In section II we consider the suggestion that ideology is pseudo-science. In section III we consider it as consisting of non-scientific views. Finally, in section IV, we draw on the preceding discussion to appraise the claim that economists' policy proposals are ideological.

I. The pursuit of scientific research is a social activity, and thus must have a sociological dimension. In an epistemological and methodological context, however, interest centres, not on the sociological aspects, but on how to appraise scientific theories. In that context any explanation, even a successful one, of how people's social position causes them to hold certain views and beliefs does not imply anything about the truth of those beliefs (Popper 1959, pp. 31–2). To see this, consider the proposition, sometimes called 'the principle of sociologism', that *all* theories are ideological. It is sometimes argued (e.g. Popper 1957, notes 7 and 8 to ch. 24, pp. 353–6) that this proposition implies the contradictory view 'all statements are false', but this may not be the case. It is sufficient to make the more modest inference that all theories are equally arbitrary. But if all theories are ideological, then so is this claim about all theories. Hence this particular theory of ideology is arbitrary, and must be rejected if the idea of objective truth is to be retained. Indeed, this is implicitly conceded when physics is deemed not to be ideological. It then follows that the socio-psychological motives which may induce people to advance certain factual views cannot imply anything about the truth of those views. To suppose the contrary is to commit the genetic fallacy, the fallacy that the truth of statements is decidable on the basis of their originators' motives in uttering them and, perhaps, believing them to be true (Rosenberg 1976, pp. 202–3). Of course, if it could be shown that economists' adherence to particular theories is conditioned by their social position, or other extraneous factors, and is unrelated to logical and empirical considerations, their methods would indeed be unscientific. Attempts to show this can be found (e.g. Wiles 1979–80), but they cannot be appraised here. Mention must be made of an idea related to, but not identical with, the

view that all theories are ideological. This view asserts that people can communicate successfully, even within a given subject such as economics, only if they share a common intellectual framework. Sir Karl Popper styled this view 'The Myth of the Framework' (1976). If this view were true, it would imply, for example, that supporters of the rational expectations, market-clearing paradigm of the functioning of a market economy could not communicate successfully with those economists who do not work within that paradigm. A glance at the professional literature shows that the view is false.

II. We saw that, if we use 'ideological' in the sense of section I, we must reject the statement, 'all statements are ideological'. This nevertheless leaves open the possibility that economics itself consists of statements which express 'biased' (i.e. false) views, although whether they *are* false is not decidable on the grounds of their originators' psychological motives, or social position. Without committing the genetic fallacy, writers who think of economics as 'impregnated with ideology' have suggested that ideological utterances be regarded as pseudo-scientific.

One suggestion is that 'ideological statements . . . be . . . defined as value judgments parading as statements of facts' (reported by Blaug 1980, p. 138), i.e. as *covert* prescriptions, all the more suspect, since they are supposedly motivated by attempts to promote some 'class interest' (Rosenberg 1976, pp. 203–4, examines this claim). It has been suggested that economics does, or must, consist only of such ideological pseudo-statements, and therefore cannot be scientific (a suggestion criticized in Klappholz 1964). No doubt a careless reader could mistake disguised value judgements for factual statements, but this possibility is a subject for psychological, rather than methodological, consideration, despite occasional suggestions to the contrary (e.g. Blaug 1980, p. 138).

Turning to statements which are descriptive, i.e. have a truth value, the following are among other suggested jointly sufficient conditions for economic statements to be ideological, i.e. pseudo-scientific: (a) that they be false; (b) that they support a given political philosophy, or be convenient for those with an interest in perpetuating some political or social order; (c) that the given political philosophy, or the convenience of the belief, be the cause of the false statements being believed (Mingat et al. 1985, pp. 353–5 and Rosenberg 1976, pp. 204–9, critically discuss these characterizations).

The philosophic problem of demarcating scientific from other kinds of discourse cannot be discussed here. It must suffice to point out that the above characterizations would render (a set of) statements pseudo-scientific if one subscribed to the epistemological view that 'true science' consists of statements known to be true by being logically derived from facts (Lakatos 1978, ch. 1). Few, if any, philosophers subscribe to this infallibilistic view of science and, in its absence, the above characterizations do not render statements pseudo-scientific (although, as noted above, (c) alone would not be a methologically satisfactory reason for an economist to support a theory). Thus, if statements are judged ideological, not because they are false, but because they are possibly false, then one could not say they are pseudo-scientific, since all scientific theories are possibly false. Again, if a universal theory is viewed as ideological because it is regarded as false, for example, as is Newton's theory, but at the same time is accepted for certain technological purposes (Klappholz and Agassi 1959, pp. 31–3), it is still not pseudo-scientific. Indeed, if such theories are regarded as pseudo-scientific, the view of ideology considered here leads to the no doubt unintended, but nevertheless absurd consequence that the available stock of pseudo-science increases with scientific progress. References to the 'convenience' of certain views, i.e. to (b), as alleged explanations of why supposedly false theories are believed, i.e. to (c), direct criticism towards individuals' conscious or unconscious motives, in

the spirit of the Sociology of Knowledge, rather than to the objective scientific issues.

So far we have discussed the possible or actual falsity of theories. Theories are falsified if observations come to light which are in conflict with them. These observations are reported in what have been called basic statements (Popper 1959, chs IV, V), i.e. statements the acceptance of which does not give rise to controversy.

For example, economists advance theories about the determinants of unemployment. These theories might be thought to be testable with the help of observations of unemployment, which, for example, lead to observation reports such as 'the level of unemployment in the UK in March 1985 was 13.3 per cent or 3.2 million people'. This is not an explanatory statement and therefore, presumably, not pseudo-scientific. However, as is well known, it is also not a basic statement, since it is controversial. Controversy is aroused, not only because the statement raises problems of statistical interpretation, but also theoretical problems, such as the observations which would be needed to measure the extent of involuntary unemployment (although, given the way unemployment is measured, large changes in the measured figures have led economists to reconsider their theories of unemployment). This is merely an example of some of the well-known problems encountered in attempts to test economic theories. Therefore, these theories are not obviously false, as seems to be required of a theory if it is to be ideological in the sense of the present Section. However, this discussion suggests that economics contains factual theories which may not be scientific, i.e. testable.

III. Factual theories which are not scientific – rather than pseudo-scientific – have been called ideological (e.g. Schumpeter 1949; Robinson 1962.) If one does not view scientific theories as consisting of statements known infallibly to be true, but rather as tentative hypotheses, which can be revised in the light of new evidence, then one can easily think of statements which are not scientific, but which nevertheless play a role in discussions of economic theories. Here we are referring to metaphysical statements, as well as to expressions of belief regarding the truth of competing theories among which do decisions can be made on the basis of tests.

Some economic theories may be testable (for example, the appearance of stagflation must be regarded as an anomaly for *all* previous economic theories that are relevant to the subject). However, many appear not to be testable. For instance, it has been held that general equilibrium theories are not testable (e.g. Hausman 1981). This consideration may account for Friedman's well-known remark that reports of the corroboration of some economic theories he endorsed are 'hard to document' (Friedman 1953, p. 22). Indeed, it has been argued that, since economic data are derived from situations which cannot be controlled for disturbing factors, statistical inference is possible only on the basis of *prior* beliefs, the differences among which cannot be objectively justified (Leamer 1983).

Where theoretical conflicts of views cannot be resolved by available evidence, it is possible to suspend judgement. However, those engaged in research need to choose a programme, that is, to judge which theory is most likely to offer the best prospects for scientific progress. This choice may be influenced by people's Weltanschauung and preferences, in short, their ideologies. In this respect the situation in economics does not seem to differ from that in other sciences, and the mere fact that ideology, in the sense of the present section, may play a part in discussions of economic theory need not give rise to 'concern for [its] conceptual status' (Rosenberg 1976, p. 202). Concern may be expressed with good reason if and when unwarranted claims to scientific knowledge are made.

IV. Historically, the charge of 'ideological bias' has been directed especially at economists' views on desirable economic policies, as suggested by remarks that economists have tended to 'justify the ways of Mammon to

men' (Robinson 1962, p. 25). We now consider this issue in the light of the preceding discussion.

It was noted above that policy recommendations cannot be deduced solely from economic theory: in addition, some value, i.e. non-scientific, premises are required. The Paretian value premise, widely adopted by economists, reflects an individualistic political philosophy and may be regarded as ideological (Klappholz 1968).

Apart from adopting the Paretian value premise, economists have advocated policies which show a preference for organizing economic activities through markets (Kearl et al. 1979). However, it is difficult to take seriously the view, referred to in sections I and II, that this stance is to be explained by those economists' 'position in the social structure' or by their 'interest in perpetuating the system'. If the preference for market-organized economic activity is less marked among, for example, sociologists, wherein lies the difference in their social position, or their interest in perpetuating the system, compared to that of economists? Thus, it is more plausible to suppose that economists' preference for markets has been shaped by the dominant paradigm of the invisible hand, and by the fact that there is the most widespread professional consensus on the consequences of overriding markets, by, for example, such policies as rent control.

It was noted that, where theoretical differences cannot be resolved, judgement may be suspended. In the case of policy, policy makers and their advisers cannot suspend judgement, since decisions cannot be avoided, even if the implicit decision is to take no action. Assuming no well-grounded consensus regarding the consequences of alternative courses of action, it seems plausible that ideological views will influence judgements on the most likely consequences, thus influencing decisions, quite apart from the value premises which are logically indispensable for reaching them. In general, there seems to be less consensus regarding the effects of policies in the area of economics than in policies based mainly on the natural sciences, although lack of consensus in the latter case is not unknown. Thus, not surprisingly, there is more scope for ideological influence in decisions about economic policy.

However, given the absence of consensus, and the relevance of economics to public policy, differences in ideological views, (be they differences in value judgements or differences in beliefs about the outcome of policies) can be viewed as part of the mechanism of the public aspect of scientific activity which promotes criticism and, through it, may help us to learn more about the issues at hand. Those for, and those against, a given policy *all* may have an ideologically based incentive to try to show, as objectively as possible, the practical consequences any given policy will have. This view is opposed to the conventional wisdom, according to which ideology is a 'Weltanschauung felt passionately *and defended unscrupulously'* (Wiles 1978–80, p. 61, italics added). Ideological views need not lead to dogmatism, or to lack of scruples, and there is, in any case, no way of ensuring the absence of dogmatic people. All one can do is to shun discussion with them.

See Also

- ▶ Philosophy and Economics
- ▶ Rhetoric of Economics
- ▶ Value Judgements

Bibliography

Blaug, M. 1980. *The methodology of economics*. Cambridge: Cambridge University Press.

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

Hausman, D.M. 1981. Are general equilibrium theories explanatory? In *Philosophy in economics*, ed. J. Pitt. Dordrecht: D. Reidel. Reprinted in *The Philosophy of economics*, ed. D.M. Hausman. Cambridge: Cambridge University Press, 1984.

Kearl, J.R., C.L. Pope, G.T. Whiting, and L.T. Wimmer. 1979. A confusion of economists. *American Economic Review* 69(2): 28–37.

Klappholz, K. 1964. Value judgments and economics. *British Journal for the Philosophy of Science*, August. Reprinted in *The philosophy of economics*, ed. D.M. Hausman. Cambridge: Cambridge University Press, 1984.

Klappholz, K. 1968. What redistribution may economists discuss? *Economica* 35(May): 194–197.

Klappholz, K., and J. Agassi. 1959. Methodological prescriptions in economics. *Economica*, February. Reprinted in *Readings in microeconomics*, ed. D.R. Kamerschen. New York: 1967.
Lakatos, 1. 1978. Introduction. In *The methodology of scientific research programmes, philosophical papers*. Vol. I, ed. G. Currie and J. Worral. Cambridge: Cambridge University Press.
Leamer, E.E. 1983. Let's take the con out of econometrics. *American Economic Review*, March. Reprinted in *Appraisal and criticism in economics*, ed. B. Caldwell. London: Allen & Unwin, 1984.
McCloskey, D.N. 1983. The rhetoric of economics. *Journal of Economic Literature*, June. Reprinted in *Appraisal and criticism in economics*, ed. B. Caldwell. London: Allen & Unwin, 1984.
Mingat, A., P. Salmon, and A. Wolfelsperger. 1985. *Methodologie économique*. Paris: Presses Universitaires de France.
Popper, K.R. 1957. *The open society and its enemies*, vol. II, 3rd ed. London: Routledge & Kegan Paul.
Popper, K.R. 1959. *The logic of scientific discovery*. London: Hutchinson.
Popper, K.R. 1976. The myth of the framework. In *The abdication of philosophy: Philosophy and the public good*, ed. E. Freeman. La Salle: Open Court.
Robinson, J. 1962. *Economic philosophy*. London: Watts. Reprinted London: Pelican Books, 1964.
Rosenberg, A. 1976. *Micro-economic laws: A philosophical analysis*. Pittsburgh: University of Pittsburgh Press.
Schumpeter, J. 1949. Science and ideology. *American Economic Review*, March. Reprinted in *The philosophy of economics*, ed. D.M. Hausman. Cambridge: Cambridge University Press, 1984.
Ward, B. 1979. *The ideal world of economics: Liberal radical and conservative economic world views*. London: Macmillan.
Wiles, P. 1979–80. Ideology, methodology, and neoclassical economics. *Journal of Post Keynesian Economics*, Winter. Reprinted in *Why economics is not a ecience*, ed. A.-S. Eichner. New York: M.E. Sharpe, 1983.

Illicit Drugs, Retail Market

Manolis Galenianos

Abstract

Three key features of the retail trade for illicit drugs are documented: moral hazard, repeated interactions and price dispersion. An interpretation of this evidence based on search and informational frictions is presented. Various policy implications of the suggested interpretation are discussed.

Keywords

Drugs; Cocaine; Heroin; Informational friction; Narcotics

JEL Classifications

J64; K14; K42

Introduction

In addition to being illegal, the retail trade for narcotics is subject to severe moral hazard. Moral hazard results from the ease with which a seller can covertly dilute ('cut') the product, while the illegality of trade prevents the emergence of institutions that solve similar informational problems in legal markets (e.g. third-party certification, product guarantees, customer reviews). Nevertheless, the market for illicit drugs does operate, albeit in a very different way from the textbook Walrasian paradigm. The purpose of this article is to review the evidence about how this market operates, to discuss ways to interpret the evidence and to consider policy against the background of the suggested interpretation.

The Evidence

This section's evidence concerns the retail trade for heroin, crack cocaine and powder cocaine. Three central features of trade will be documented: moral hazard, repeated interactions and price dispersion.

The first feature is that retail transactions for illegal drugs are subject to moral hazard. Table 1, taken from Galenianos et al. (2009), is based on the **STRIDE** dataset of undercover Drug Enforcement Administration purchases and documents an extreme instance of the moral hazard – the *rip-off*, a transaction in which the buyer is sold essentially zero-purity drugs. A significant fraction of 'street-

Illicit Drugs, Retail Market, Table 1 Rip-offs in trades with value <= $100 in 1983 dollars

Drug	Average purity	Percentage of all trades that are rip-offs (i.e., <= 2% purity)	Average price of rip-offs (std. dev. of price)	Average price of non rip-offs (std. dev. of price)
Heroin	31%	10.3%	$53 (22.8)	$57 (20.6)
$N = 12{,}716$				
Crack	68%	7.8%	$32 (21.3)	$38 (24.6)
$N = 16{,}202$				
Cocaine	54%	5.1%	$35 (21.8)	$53 (25.8)
$N = 5{,}362$				

level' transactions are seen to be total rip-offs. Most important, the price paid in a rip-off is not appreciably different from that of a non-rip-off transaction, suggesting that buyers cannot observe dilution.

The practice of selling drugs in branded bags ('dope stamps') is further corroborating evidence of a quality problem in the illicit drugs trade. Wendel and Curtis (2000) describe the usage of dope stamps in New York City in a very interesting ethnographic study. The purported effect of a dope stamp is quality certification. The stamps could be boasts of quality ('America's Choice', 'Dynamite'), status brands ('Dom Perignon', 'Gucci') or even corporate names ('Exxon'). However, because they can be faked by 'unscrupulous' competitors, the certification value of a dope stamp is limited and short-lived (a couple of days, often). It therefore seems clear that dope stamps do not solve the quality certification problem.

The prevalence of long-term relationships is the second key feature of the drugs trade. The Table 2 provides direct evidence as to the prevalence of repeated interactions from the ADAM dataset. The data is based on voluntary interviews with a random sample of arrestees who self-report their drug habits. Conditional on purchasing drugs, the respondents report engaging in multiple transactions during the previous month, but use only a small number of suppliers.

The process by which buyers manage, over time, to hook up with a seller with whom they develop a long-term relationship is described in the ethnographic study by Hoffer (2005). However, not all sellers need have repeat business. The ethnographic literature reports on sellers who specialise in selling rip-offs. Hamid (1992) refers to these sellers as 'zoomers', a street expression due to the practice of selling bogus drugs and then disappearing.

Illicit Drugs, Retail Market, Table 2 Repeated transactions

Previous month	Heroin $N = 3{,}249$	Crack $N = 8{,}321$	Cocaine $N = 4{,}302$
Average number of purchases	18.7	13.2	7.2
Average number of suppliers used	2.7	3.3	1.8

The market's third feature is the very substantial dispersion in the price/quality ratio. Price dispersion in the drugs market is documented in Reuter and Caulkins (2004), where it is shown to be several times higher than that observed in markets for licit goods. Table 3 (reproduced from Galenianos et al. 2009) shows that there is substantial variation in the amount of pure drugs that can be had for $100. Additionally, it shows that dispersion occurs mostly within a location and time unit and hence is not due to time or local price variation.

Models and Policy

There is an extensive literature that focuses on the *demand* for illicit drugs, discussing the role of harmful addiction, rationality and discounting (Becker and Murphy 1988). Formal theoretical models of the *market structure* are tied to

Illicit Drugs, Retail Market, Table 3 Price dispersion

Drug	Sample cities with => 400 obs.	Mean pure grams per $100	Standard deviation	Coefficient of variation
Heroin	Full	0.38	0.58	1.45
$N = 19{,}072$	w/o city-year fixed effect	0	0.47	1.24
Crack	Full	1.73	1.49	0.86
$N = 20{,}262$	w/o city-year fixed effect	0	1.32	0.76
Cocaine	Full	2.14	1.98	0.93
$N = 18{,}862$	w/o city-year fixed effect	0	1.59	0.74

traditional economic assumptions of perfect information and centralised markets (e.g. Becker et al. 2006). Within that framework, all types of enforcement at all levels of the supply chain are generally lumped together and modelled as a 'cost of doing business' for the dealer. However, these assumptions abstract from important features of the market by ignoring the choice of purity (moral hazard) and search costs.

An alternative approach is to explicitly model the search and informational frictions that are present in this market. Galenianos et al. (2009) develop an equilibrium search model of repeated trade with unobservable quality. In that model consumers of drugs engage in costly search for sellers and the level of purity is chosen by each seller. A key assumption is that buyers can only determine the quality of drugs *after* the trade is consummated, which distinguishes this model from Burdett and Mortensen (1998), its counterpart from the labour search literature. The focus of the analysis is on determining the level of quality that will be traded for a given amount of money, that is, the *affordability* of (high-quality) drugs in equilibrium.

The informational frictions lead to severe quality problems, putting the market at risk of collapse. Indeed, the incentives for opportunistic behaviour by sellers are only mitigated by the possibility of forming long-term relationships with buyers: a seller who wants to keep a customer will not rip him off and, as a result, moral hazard does not necessarily foreclose the possibility of trade. In equilibrium, some sellers will offer good quality to increase their sales; others, however, will specialise in ripping off their customers. Introducing moral hazard, therefore, goes a long way towards accounting for the key stylised facts presented above: the mass of sellers who cheat their customers by providing zero-purity drugs, the importance of long-term relationships and the wide dispersion in the price/quality ratio.

In evaluating policy, the conventional view is rather generic: tougher penalties and more law enforcement, at any level of the supply chain, should help reduce the affordability of drugs. In fact, there is little evidence that recent efforts to increase penalties and law enforcement have measurably reduced the availability of drugs. The price of a pure gram of cocaine or heroin has declined substantially during the periods when budgets on law enforcement rose and penalties increased (Caulkins et al. 2004). In Galenianos et al. (2009) different enforcement instruments can impact the retail affordability of drugs in complex and sometimes counterintuitive ways. For example, to the extent that police enforcement makes it more risky for buyers to search for new sellers, the long-term relationship between buyers and sellers is strengthened, which in turn alleviates moral hazard and expands the possibility of trade. These findings highlight the need for developing models that are tailored to the characteristics of the market before evaluating policy.

Finally, at a somewhat more speculative level, the analysis in Galenianos et al. (2009) suggests alternative channels to suppress the market. If it is

true that the market is undermined by moral hazard then economic theory suggests *leveraging the moral hazard*, i.e. inducing sellers to *dilute more*. Within the model, a policy of reducing the sentences of sellers who sell low-purity drugs leads to an increase in 'cheating' and hence an increase in the average price of a pure gram of drugs. In addition, this is accomplished by actually reducing incarceration rates relative to current levels, thus simultaneously achieving two seemingly contradictory desiderata.

See Also

- ▶ Addiction
- ▶ Moral Hazard
- ▶ Search Theory
- ▶ Search Theory (New Perspectives)

Bibliography

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

Becker, G.S., K.M. Murphy, and M. Grossman. 2006. The market for illegal goods: The case of drugs. *Journal of Political Economy* 114(1): 38–60.

Burdett, K., and D.T. Mortensen. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39: 257–273.

Caulkins, J., R.L. Pacula, J. Arkes, P. Reuter, S. Paddock, M. Iguchi, and J. Riley. 2004. *The price and purity of illicit drugs: 1981 through the second quarter of 2003*. Washington, DC: Office of National Drug Control Policy, The White House.

Galenianos, M., R. Pacula, and N. Persico. 2009. A search-theoretic model of the retail market for illicit drugs. *NBER WP 14980*.

Hamid, A. 1992. The developmental cycle of a drugs epidemic: The cocaine smoking epidemic of 1981–1991. *Journal of Psychoactive Drugs* 24(4): 337–348.

Hoffer, L.D. 2005. *Junkie business: The evolution and operation of a heroin dealing network*. Florence: Wadsworth Publishing.

Reuter, P., and J. Caulkins. 2004. Illegal lemons: Price dispersion in the cocaine and heroin markets. *UN Bulletin on Narcotics* 56: 141–165.

Wendel, T., and R. Curtis. 2000. The heraldry of heroin: 'Dope stamps' and the dynamics of drug markets in New York City. *Journal of Drugs Issues* 30(2): 225–260.

Immigration and the City

Anna Hardman

Abstract

At the end of the 20th century, international migrants, legal and undocumented, were a highly visible and economically significant feature of major cities in high- and middle-income countries, including the United States. As numbers of immigrants rose, many were concentrated spatially in a small number of cities ('ports of entry') and within those cities in ethnically homogeneous neighbourhoods, enclaves or ghettos. An extensive literature documents the impact of immigrants on host cities, examines their patterns of assimilation and explores their interactions with native-born populations and previous immigrants.

Keywords

Economies of agglomeration; Ghettoes; Housing markets; Information sharing; Internal migration; International migration; Labour market discrimination; Migrants in cities; Neighbourhoods; Network externalities; Residential assimilation; Residential segregation; Selection bias and self-selection; Spatial correlation; Urban agglomeration; Urban economics

JEL Classifications

R23

At the end of the 20th century, international migrants, legal and undocumented, were again a highly visible and economically significant feature of major cities in high- and middle-income countries, including the United States. As numbers of immigrants rose, many were concentrated spatially in a small number of cities ('ports of entry') and within those cities in ethnically homogeneous neighbourhoods (enclaves or ghettos).

That was not a new phenomenon: in the 'first great migration' to the United States in the late 19th and early 20th centuries immigrants were highly concentrated and a highly visible feature of the largest cities. In 1870, the foreign-born constituted 35.6 per cent of the population of US cities over 100,000 and almost 50 per cent of the population of San Francisco and Chicago, though only 14.4 per cent of the national population. By 1940 the immigrant share had declined to 16.2 per cent of the population of cities over 100,000 and 8.8 per cent of the total population (Gibson and Lennon 1999, Tables 18 and 23).

In 15 Organisation for Economic Co-operation and Development (OECD) countries at the beginning of the 21st century the foreign-born made up between 8.3 and 32.6 per cent of the national population (Dumont and Lemaitre 2005, Table 1). In the United States in 2000, the foreign-born constituted 26.9 per cent of the population in the central cities of metropolitan areas with a population of five million or more and 16.2 per cent in the suburbs. In the cities of New York and Los Angeles the foreign-born made up 35.9 and 40.9 per cent of the population respectively. In the ten metropolitan areas with the largest immigrant populations the foreign-born were between 35 and 54.9 per cent of the population. The foreign-born were correspondingly rare outside large cities: less than four per cent of the population in metropolitan areas with a population of 500,000 or less and even rarer outside metropolitan areas (US Census of Population 2000). Similarly in the UK in 2001 8.3 per cent of the total population was born overseas. In the same year, the foreign- born were about 25 per cent of the London metropolitan area's total population, and were concentrated in a few neighbourhoods. For example, in Southall, Wembley, Hyde Park and Kensington, over 45 per cent of the population was foreign-born (National Statistics 2005; BBC News 2007).

As a result, whereas until the 1970s race and ethnicity were typically absent from analyses of urban economies (in Europe) or modelled as a black–white dichotomy (in the United States), by the mid-1980s economists had begun to explore the impact of immigrants from a wide range of source countries on cities beyond their effect on the wages and employment of natives. Urban economists have explored residential assimilation, looking at location choices, crowding and housing tenure and asked whether the location and housing consumption of immigrants relative to natives has differed because of selection, country of origin or changing make-up of successive cohorts of immigrants. The literature is dominated by studies of the United States both because of its rapidly growing immigrant population and because of micro (individual or household-level) and spatially disaggregated data on immigrant status, race and ancestry.

Immigrants are attracted to ports of entry or places with a stock of previous immigrants, because migration is path-dependent, because immigrants in enclaves benefit from *network externalities* and because immigrant enclaves offer *economies of agglomeration.* As a result, immigrants and particularly unskilled immigrants are less mobile within host societies than the native-born. The behaviour of the native- born in the host economy also drives spatial outcomes, both because of *discrimination* or avoidance of immigrants in labour and housing markets and because natives' location decisions across cities within a host country are more sensitive to wages than those of immigrants. There is evidence that the concentration of immigrants in ports of entry has led some US natives to leave gateway cities or to move to alternative destinations (Filer 1992).

Early empirical work on immigration and wages estimated the impact of immigration on the labour market using a cross-sectional 'spatial correlations' approach that compared wages over time in metropolitan areas with different proportions of immigrant stocks and flows. The spatial correlations approach generally found weak links at best between the immigrant share and the wages and on employment of natives, both in the USA (Borjas 1994) and more recently in the UK (Hatton and Tani 2005). If natives' location decisions are more sensitive to labour conditions than immigrants', then the observed wage impact of immigration is attenuated because it is dispersed across the whole economy rather than concentrated in the port of entry cities.

The impact of immigration on internal migration has been pursued by geographers and demographers (Wright et al. 1997; Kritz and Gurak 2001) with some expressing fear that the United States faced demographic or spatial 'balkanization' or the concentration of immigrants in a few cities shunned by natives (Frey 1995, 1996). However, in the United States immigrants' location patterns are changing. In the 1990s growing numbers of immigrants moved to urban and suburban areas remote from the traditional ports of entry. The immigrant population grew more rapidly in non-traditional destinations. For example, the US 2000 Census found that in ten metropolitan areas (with a median population of over 160,000) over two-thirds of the foreign-born population had entered the USA in the previous decade. The new immigrants were moving to metropolitan areas where only a median 4.55 per cent of the population was foreign-born by 2000. Some were in states without a recent tradition of immigration (Iowa, Indiana, North and South Dakota and Nebraska); others were in or close to states with a significant immigrant presence already (Arizona, Georgia, North Carolina and Tennessee).

A notable feature since 1985 is the increasing dispersion of Mexican immigrants, who for a long time were highly concentrated in Los Angeles and elsewhere in Southern California and Texas (Alba et al. 1999). That migration is also credited with changing the industry mix in destination regions (Card and Lewis 2007). Immigrants who move again within the USA have higher skills than other new immigrants. Moreover, migration beyond immigrant gateways and enclaves is associated with faster assimilation, although this is in part probably attributable to reverse causation since secondary migrants are self-selected (Zhang 2004, 2006).

In the absence of detailed information on the immediate spatial areas where immigrants live, most of the analysis of residential location, however, focuses on individuals. Immigrants often live in households with partners or family members who are natives or second- or third-generation immigrants. Household-level analysis of confidential Current Population Survey data for Los Angeles shows much greater dispersion of immigrants living in mixed households (Ellis and Wright 2005).

The urban economics literature on immigrants in cities has been concerned with 'residential assimilation': the progress of new immigrants towards parity with natives in housing tenure, consumption of housing and intra-city location, in or outside of ethnic enclaves (see, for example, Painter et al. 2001). While US immigrant homeownership rates are consistently lower than natives', they rise with age and years. Increases in the gap between natives and immigrants in homeownership rates between 1980 and 2000 are explained by differences in location decisions and by changes in the national origin mix of the immigrant population that are associated with lower skills and wages for the most recent immigrant waves (Borjas 2002).

In contrast to labour markets, where impacts of immigration on wages have been elusive, there is evidence that the arrival of immigrants raises metropolitan area housing prices and rents (Saiz 2003). Saiz and Wachter (2006) also find immigration associated with relatively slower house price appreciation in immigrant enclaves. The latter is attributed both to native avoidance and to low-income immigrants' preference for the cheapest housing.

Another facet of housing consumption and hence residential assimilation is residential 'crowding' (large numbers of occupants per dwelling or per room). Crowding increased in the USA in the 1980s and the 1990s, after decreases in every decade from 1940 to 1980; the increases were almost all in areas with large concentrations of immigrants. Cohort studies have found that immigrants initially choose higher densities, which decline with time in the United States for most ethnic groups with the exception of Hispanic immigrants (Myers and Lee 1996; Simmons 2002).

Economists have begun to explore the role of ethnic enclaves, neighbourhoods with a high concentration of immigrants, usually from the same source country or region. They are characterized by forces that parallel those that drive the formation of cities and concentrations of firms: shared

inputs (enclaves offer stores which provide ethnic foods, clothing, goods used both for consumption and for production by local firms); information-sharing as immigrants in enclaves benefit from news of job opportunities and learn skills essential in the job market and for everyday life in the host country; lastly, new and particularly unskilled immigrants as well as entrepreneurs in the enclave benefit from labour-market pooling in the enclave labour market.

Empirical studies of enclave economies provide evidence that immigrants value location near others from the same source country or region and that there are measurable economic benefits. Gonzalez (1998) estimated the implicit 'price of culture' using 1990 Census data for California and Texas, and found both lower earnings and higher rents for Mexican immigrants in enclaves with larger concentrations of Mexicans. Other studies find that immigrants within enclaves earn less than those outside, but a problem with such studies is that immigrants in enclaves are self-selected. A notable recent finding comes from Edin et al. (2003) who exploit a natural experiment in Sweden in which asylum- seekers and refugees were randomly assigned to different cities. They find evidence that selection bias leads to significant underestimates of the value of living in enclaves, with an earnings gain in the order of four to five per cent for migrants living in enclaves, compared with the earnings losses observed before correcting for selection.

See Also

- ▶ Ghettoes
- ▶ Housing Supply
- ▶ Urban Agglomeration

Bibliography

Alba, R., J.R. Logan, B.J. Stults, G. Marzan, and W. Zhang. 1999. Immigrant groups in the suburbs: A reexamination of suburbanization and spatial assimilation. *American Sociological Review* 64: 446–460.

BBC News. 2007. Born abroad: An immigration map of Britain. Online. Available at: http://news.bbc.co.uk/2/shared/spl/hi/uk/05/born_abroad/around_britain/html/overview.stm. Accessed 30 Jan 2007.

Borjas, G.J. 1994. The economics of immigration. *Journal of Economic Literature* 32: 1667–1717.

Borjas, G.J. 2002. Homeownership in the immigrant population. *Journal of Urban Economics* 52: 448–476.

Card, D., and E. Lewis. 2007. The diffusion of Mexican immigrants during the 1990s: Explanations and impacts. In *Mexican immigration to the United States*, ed. G. Borjas. Chicago: University of Chicago Press and NBER.

Dumont, J.-C., and G. Lemaitre. 2005. Counting immigrants and expatriates in OECD countries: A new perspective. Social, employment and migration working papers no. 25. Paris: OECD.

Edin, P.-A., P. Fredriksson, and O. Åslund. 2003. Ethnic enclaves and the economic success of immigrants – Evidence from a natural experiment. *Quarterly Journal of Economics* 118: 329–357.

Ellis, M., and R. Wright. 2005. Assimilation and differences between the settlement patterns of individual immigrants and immigrant households. *Proceedings of the National Academy of Sciences* 102: 15325–15330.

Filer, R.K. 1992. The effect of immigrant arrivals on migratory patterns of native workers. In *Immigration and the work force: Economic consequences for the United States and source areas*, ed. G.J. Borjas and R.B. Freeman. Chicago: University of Chicago Press.

Frey, W.H. 1995. Immigration and internal migration flight from US metropolitan areas: Toward a new demographic balkanisation. *Urban Studies* 32: 733–757.

Frey, W.H. 1996. Immigration, domestic migration, and demographic balkanization in America: New evidence for the 1990s. *Population and Development Review* 22: 741–763.

Gibson, C.J., and E. Lennon. 1999. Historical census statistics on the foreign-born population of the United States: 1850–1990. Working paper no. 29. Washington, DC: Population Division, US Bureau of the Census.

Gibson, C.J., and K. Jung. 2006. Historical census statistics on the foreign-born population of the United States: 1850–2000. Population division working paper no. 81. Washington, DC: US Bureau of the Census.

Gonzalez, A. 1998. Mexican enclaves and the price of culture. *Journal of Urban Economics* 43: 273–291.

Hatton, T.J., and M. Tani. 2005. Immigration and inter-regional mobility in the UK 1982–2000. *Economic Journal* 115: 342–358.

Kritz, M.M., and D.T. Gurak. 2001. The impact of immigration on the internal migration of natives and immigrants. *Demography* 38: 133–145.

Myers, D. 1999. Immigration: Fundamental force in the American city. *Housing Facts and Findings* 1(4): 3–5. Fannie Mae Foundation.

Myers, D., and S.W. Lee. 1996. Immigration cohorts and residential overcrowding in southern California. *Demography* 33: 51–65.
National Statistics. 2005. People and migration: Foreign-born. Online. Available at: http://www.statistics.gov.uk/cci/nugget.asp7id = 1312. Accessed 1 Feb 2007.
Painter, G., S. Gabriel, and D. Myers. 2001. Race, immigrant status and housing tenure choice. *Journal of Urban Economics* 49: 150–167.
Saiz, A. 2003. Room in the kitchen for the melting pot: Immigration and rental prices. *Review of Economics and Statistics* 85: 502–521.
Saiz, A., and S. Wachter. 2006. Immigration and the neighborhood. Working paper 06–22, Research Department, Federal Reserve Bank of Philadelphia.
Simmons, P.A. 2002. Patterns and trends in overcrowded housing: Early results from Census 2000. Fannie Mae Foundation Census Note 09. Fannie Mae Foundation. Online. Available at: http://www.fanniemaefoundation.org/programs/pdf/census/census_note9.pdf. Accessed 1 Feb 2007.
US Census of Population. 2000. United States – Metropolitan areas. SF 3 Table GCT-P10. Online. Available at: http://factfinder.census.gov. Accessed 2 Apr 2007.
Wright, R.A., M. Ellis, and M. Reibel. 1997. The linkage between immigration and internal migration in large metropolitan areas in the United States. *Economic Geography* 73: 234–254.
Zhang, W. 2004. Secondary migration of recent immigrants in the United States: A study of Mexicans and Chinese. Ph.D. thesis, State University of New York at Albany.
Zhang, W.C. 2006. Internal migration of Mexicans in the United States, 1990 and 2000. Brown University Department of Sociology. Presented at the Annual Meeting of the Population Association of America, Los Angeles, 30 March–1 April.

Immiserizing Growth

Jagdish N. Bhagwati

Keywords

Autarky; Directly Unproductive Profit-seeking (DUP) activities; Distortions; Free trade; Immiserizing growth; Shadow pricing; Tariffs; Tariff seeking; Terms of trade

JEL Classifications

F1

The theory of immiserizing growth has been developed by theorists of international trade, though it has recently been the focal point of research also by mathematical economists. It is central to understanding several important paradoxes in economic theory and has significant policy implications.

That growth in a country could immiserize it is a paradox that was first noted by trade theorists such as Bhagwati (1958) and Johnson (1955) in the context of the post-war discussions of dollar shortage. They established conditions under which, in a two-country, two-traded-goods framework of conventional theory, the growth-induced deterioration in the terms of trade would outweigh the primary gain from growth. It was shown that this paradox, unlike the paradox of donor-enriching and recipient-immiserizing transfers, was compatible with Walras-stability.

The phrase 'immiserizing growth' was invented by Bhagwati (1958) and has now been widely accepted (including by literary editors who have long ceased to insist on changing it to the correct English versions such as 'immiserating'), the theory itself being generally attributed (for example, Johnson 1967) to this 1958 article. Interestingly, as often in economics, Bhagwati happened to chance upon an early contribution by Edgeworth (1894), where Edgeworth developed an example of what he called 'indamnifying' growth; and the controversy surrounding this result at the time and its relationship to the Bhagwati–Johnson analyses of the 1950s was reviewed in Bhagwati and Johnson (1960).

Later, Johnson (1967) demonstrated another paradox of immiserizing growth. If a small country had a distortionary tariff in place, and then exogenously it experienced growth, the result again could be to immiserize the country. Later, Bertrand and Flatters (1971) and Martin (1977) established formally the conditions under which this new paradox of immiserizing growth could arise.

Bhagwati (1968) got to the bottom of these paradoxes and produced the central insight that explains why these, and other immiserizing-growth paradoxes, can readily arise. He showed that, if an economy was suboptimally organized, the primary gain from growth, measured

hypothetically as if the economy had an optimal policy in place before and after the growth, could be outweighed by accentuation of the loss from the distortion-induced suboptimality when growth occurred. In the original Bhagwati (1958) example, since the terms of trade could deteriorate, the economy had monopoly power in trade but was following free trade policy which is evidently suboptimal. In the Johnson (1967) example, the tariff was being used by a small country with given terms of trade and was therefore also a suboptimal policy. In both cases the suboptimal policy produced losses which were accentuated by the growth and then managed to outweigh the primary gains from growth that would have occurred if optimal policies were in place. The result was a powerful generalization that placed the theory of immiserizing growth squarely into the central theory of distortions and policy intervention (Srinivasan 1987) that lies at the core of the modern theory of trade and welfare. Evidently, immiserizing-growth paradoxes could arise only if there was a distortion present.

This central result has immediate implications. If an economy has a suboptimal money supply, growth could be immiserizing. If trade policy is highly distorted, growth could be immiserizing. The well-known results of trade theory, which show that free trade need not be welfare-improving relative to autarky (for example, Haberler 1950) under distortions are also seen as instances of immiserizing-growth theory; free trade augments the availability set relative to autarky, implying 'as-if' growth, and if distortions are present, then there is no surprise to the immiseration that free trade brings. Again, if a country uses tariffs to induce foreign investment (the so-called tariff-jumping investment that developing countries often used in the post-war period), such investment could immiserize the host country: this being a simple extension of the Johnson (1967) demonstration, argued to be relevant to analysis of developing countries in Bhagwati (1978), and analysed extensively in Bhagwati (1973), Brecher and Alejandro (1977), Hamada (1974), Minabe (1974), Uzawa (1969) and Brecher and Findlay (1983). Yet another important insight from the immiserizing-growth theory is that, in the new and growing theory of DUP (directly-unproductive profit-seeking) activities, which incorporates several quasi-political activities essentially into the corpus of economic theory, a DUP activity that wastes resources directly need not cause ultimate loss of welfare. This is because the waste may occur from a suboptimal situation, thus resulting in welfare-improvement paradoxically. This is the obverse of immiserizing growth: in one case, growth immiserizes; in the other, throwing away or wasting resources enriches. This is at the heart of the contention in Bhagwati (1980) that an exogenous tariff at t per cent may be welfare-superior to an endogenous tariff, procured by tariff-seeking lobbies that have diverted uses to such DUP activity, also at t per cent. Several such implications of the theory of immiserizing growth are discussed in Bhagwati and Srinivasan (1983, ch. 25).

Two further developments need to be cited. First, the dual of immiserizing growth, when such growth is due to factor accumulation, clearly yields negative shadow factor prices. This aspect is relevant to certain formulations in cost–benefit analysis; see, in particular, Findlay and Wellisz (1976), Diamond and Mirrlees (1976), Srinivasan and Bhagwati (1978), Bhagwati et al. (1978) and Mussa (1979).

Next, mathematical economists such as Aumann and Peleg (1974), and then Mas-Colell (1976) and Mantel (1984) among others, have rediscovered the original immiserizing-growth paradox, illustrating how economists working apart or in different traditions may rediscover one another's findings, often decades apart. A synthesis of the two literatures has been provided in Bhagwati et al. (1984). A complete and formal reconciliation of the conditions established in Bhagwati (1958) and in Mas-Colell (1976) and Mantel (1984) for the original immiserizing-growth paradox is provided by Hatta (1984).

See Also

- ▶ Directly Unproductive Profit-Seeking (DUP) Activities
- ▶ Terms of Trade

Bibliography

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

Bertrand, T., and F. Flatters. 1971. Tariffs, capital accumulation and immiserizing growth. *Journal of International Economics* 1: 453–460.

Bhagwati, J. 1958. Immiserizing growth: A geometrical note. *Review of Economic Studies* 25: 201–205. Reprinted in *International trade*: *selected readings*, ed. J. Bhagwati, Cambridge, MA: MIT Press, 1981.

Bhagwati, J. 1968. Distortions and immiserizing growth: A generalization. *Review of Economic Studies* 35: 481–485. Reprinted in *The theory of commercial policy*, ed. J. Bhagwati, vol. 1. Cambridge, MA: MIT Press, 1983.

Bhagwati, J. 1973. The theory of immiserizing growth: further applications. In *International trade and money*, ed. M. Connolly and A. Swoboda. Toronto: University of Toronto Press.

Bhagwati, J. 1978. *Foreign trade regimes and economic development: The anatomy and consequences of exchange control*. Cambridge, MA: Ballinger.

Bhagwati, J. 1980. Lobbying and welfare. *Journal of Public Economics* 14: 355–363.

Bhagwati, J., and H.G. Johnson. 1960. Notes on some controversies in the theory of international trade. *Economic Journal* 60: 74–93.

Bhagwati, J., and T.N. Srinivasan. 1983. *Lectures on international trade*. Cambridge, MA: MIT Press.

Bhagwati, J., T.N. Srinivasan, and H. Wan Jr. 1978. Value subtracted, negative shadow prices of factors in project evaluation, and immiserizing growth: Three paradoxes in the presence of trade distortions. *Economic Journal* 88: 121–125.

Bhagwati, J., R. Brecher, and T. Hatta. 1984. The paradoxes of immiserizing growth and donor-enriching 'recipient-immiserizing' transfers: A tale of two literatures. *Weltwirtschaftliches Archiv* 120 (4): 228–243.

Brecher, R., and C. Diaz-Alejandro 1977. Tariffs, foreign capital and immiserizing growth. *Journal of International Economics* 7: 317–322. Reprinted in *International trade*: *Selected readings*, ed. J. Bhagwati, Cambridge, MA: MIT Press, 1981.

Brecher, R., and R. Findlay. 1983. Tariffs, foreign capital and national welfare with sector-specific factors. *Journal of International Economics* 14: 277–288.

Diamond, P., and J. Mirrlees. 1976. Private constant returns and public shadow prices. *Review of Economic Studies* 43: 41–48.

Edgeworth, F.Y. 1894. The theory of International values. *Economic Journal* 4: 35–50, 424–443, 606–638.

Findlay, R., and S. Wellisz. 1976. Project evaluation, shadow prices and trade policy. *Journal of Political Economy* 84: 543–552.

Haberler, G. 1950. Some problems in the pure theory of international trade. *Economic Journal* 60: 223–240.

Hamada, K. 1974. An economic analysis of the duty-free zone. *Journal of International Economics* 4: 225–241.

Hatta, T. 1984. Immiserizing growth in a many-economy setting. *Journal of International Economics* 17: 335–345.

Johnson, H.G. 1955. Economic expansion and international trade. *Manchester School of Economic and Social Studies* 23 (2): 95–112.

Johnson, H.G. 1967. The possibility of income losses from increased efficiency or factor accumulation in the presence of tariffs. *Economic Journal* 77: 151–154. Reprinted in *International trade*: *Selected readings*, ed. J. Bhagwati. Cambridge, MA: MIT Press, 1981.

Mantel, R. 1984. Substitutability and the welfare effects of endowment increases. *Journal of International Economics* 17: 325–334.

Martin, R. 1977. Immiserizing growth for a tariff-distorted, small economy. *Journal of International Economics* 3: 323–326.

Mas-Colell, A. 1976. En torno a una propiedad poco atractiva del equilibrio competitivo. *Moneda y Credito* 136: 11–27.

Minabe, N. 1974. Capital and technology movements and economic welfare. *American Economic Review* 64: 1088–1100.

Mussa, M. 1979. The two-sector model in terms of its dual: A geometric exposition. *Journal of International Economics* 9: 513–526. Reprinted in *International trade: Selected readings*, ed. J. Bhagwati. Cambridge, MA: MIT Press, 1981.

Srinivasan, T.N., and J. Bhagwati 1978. Shadow prices for project selection in the presence of distortions: Effective rates of protection and domestic resource costs. *Journal of Political Economy* 86: 97–116. Reprinted in *International trade*: *Selected readings*, ed. J. Bhagwati. Cambridge, MA: MIT Press, 1981.

Uzawa, H. 1969. Shinon jiyutato kokumin keizai [Liberalization of foreign investments and the national economy]. *Ekonomisuto* 23: 106–122 (in Japanese).

Impatience

Larry G. Epstein

Impatience refers to the preference for earlier rather than later consumption an idea which stems from Böhm-Bawerk (1912) and Fisher (1930), among others. Preference orderings that exhibit impatience are also described as being myopic or as embodying discounting. Because in many contexts the future has no natural

termination date, an infinite horizon framework is most appropriate and convenient for the analysis of many problems in intertemporal economics. The open-endedness of the future raises several issues surrounding impatience (its presence, degree, and the precise form it takes) which do not arise in finite horizon models.

Consider a world with a countable infinity of time periods or generations, $t = 0, 1, \ldots, T, \ldots$, where there is a single good which can be consumed or accumulated. Let $x = (x_0, \ldots, x_t, \ldots)$ represent a consumption programme where x_t denotes the consumption of the representative consumer for the tth generation. Given an initial (capital) stock k_0 of the good, and a technology that transforms capital into a flow output, the set of feasible consumption programmes, denoted $S(k_0)$, is determined.

At issue is the optimal programme of consumption and accumulation. Suppose it is determined by a central planner who ranks programmes in $S(k_0)$ according to the utility functional

$$U(x) = \sum_0^\infty (1+\rho)^{-t} u(x_t). \tag{1}$$

This is a common specification. For $\rho = 0$ it dates from Ramsey (1928); for the general case see Koopmans (1966). The instantaneous utility function $u(\cdot)$ is increasing and concave (diminishing marginal utility).

The parameter ρ equals the rate of time preference. Impatience (in the sense of any of the precise definitions given below) is present if (and only if) $\rho > 0$. There is a preliminary technical problem with (1) for some values of ρ. When $\rho = 0$, for example, the infinite sum in (1) diverges for many of the paths to be compared. Ramsey provides one device for getting around this difficulty. Another device is von Weizsäcker's (1965) overtaking criterion, according to which x^* is *optimal* in $S(k_0)$ if it is feasible and if for any other feasible path x,

$$\sum_0^T (1+\rho)^{-t} u(x_t^*) \geq \sum_0^T (1+\rho)^{-t} u(x_t),$$

for all sufficiently large T. This notion of optimality is welldefined for any value of ρ, even for negative values; and an optimal x^* maximizes U on $S(k_0)$ if $U(x^*)$ is finite.

The specification of ρ is crucial and presumably reflects the ethical principles of the planner. Ramsey (1928) objects to discounting on ethical grounds and thus assumes $\rho = 0$. But Koopmans (1966, 1967) argues that there are technical limitations on the specification of ρ which are imposed by the requirement that an optimal plan x^* exist for a range of choice environments. The potential difficulty is readily understood: a positive return to saving provides an incentive to postpone consumption. Positive (negative) discounting provides an offsetting (reinforcing) incentive. Finally, diminishing marginal utility and diminishing marginal productivity in production induce a smoothing of consumption over time. For many specifications, the net incentive is to postpone and to do so idefinitely, which is clearly not optimal. Consequently an optimal programme fails to exist. The existence problem is mitigated the larger is ρ, in the sense that if $\rho_1 < \rho_2$ and if an optimum in $S(k_0)$ exists when $\rho = \rho_1$, then it exists also when $\rho = \rho_2$. In particular, in order that an optimum exist in several simplified but commonly specified choice environments, it is necessary that $\rho > 0$ and hence that the future be discounted. (See also von Weizsäcker 1965.)

The existence of solutions to optimization problems is a basic question in mathematical programming which is most commonly resolved by application of the Weierstrass Theorem (or its many extensions). The Theorem guarantees existence of a solution if the objective function is continuous and the constraint set is compact. It is valid in general topological spaces and so is applicable also to the present setting where the choice variable x lies in an infinite dimensional space. The Theorem is the basis for the proof by Magill (1981) of the existence of an optimum to infinite horizon optimization problems. When specialized to the constant discount rate functional (1), his analysis confirms the consequences for existence of large ρ. Moreover, it shows 'why' a large ρ is beneficial – the larger is ρ, the more stringent the

form of continuity satisfied by the utility functional and hence the broader the class of constraint sets to which the Weierstrass Theorem is applicable.

To pursue the link between impatience and continuity, it is necessary to consider the latter more carefully. First, however, restrict attention to bounded consumption profiles, that is, to the set

$$L_+^\infty = \{x = (x_0, \ldots, x_t, \ldots) : \quad x_t \geq 0 \text{ for all } t \text{ and } \sup x_t < \infty\}.$$

Secondly, the existence of a utility function is an unnecessarily restrictive assumption. Thus consider preference relations $\succcurlyeq$ on L_+^∞, with strict preference denoted by $\succ$.

To discuss continuity, we need to specify a topology for L_+^∞; that is, we need to define what it means for two consumption paths to be 'close' to one another. This is most simply done by specifying when a sequence of consumption paths $\{x^n = (x_0^n, x_1^n, \ldots, x_t^n, \ldots)\}_{n=1}^\infty$ *converges* to a path x in L_+^∞. (Strictly speaking, generalized sequences called nets should be used, but the use of sequences is adequate for this informal discussion.) *For* many topologies that are of interest in economics 'closeness' can be measured by a *metric* or distance function d such that $d(x, y)$ measures the 'distance' between x and y. When such a metric exists, convergence of $\{x^n\}$ to x means simply that $d(x^n, x)$ approaches 0 as $n \to \infty$, in which case we refer to the *d-convergence* of the sequence.

Table 1 defines four topologies by specifying the conditions for convergence imposed by each. When a metric exists, it is also specified. Of course many other plausible topologies could be considered.

Continuity of a preference relation means roughly that consumption paths that are close to one another are ranked similarly vis-a-vis other paths. More formally, say that the relation $\succcurlyeq$ is continuous in the topology Γ (or Γ-continuous) if for each x and y in $L^\infty{}_+$, and for any sequences $\{x^n\}$ and $\{y^n\}$ that converge to x and y respectively according to Γ, it is the case that

$$x \succ y \Rightarrow x \succ y^n \quad \text{and} \quad x^n \succ y$$

for all sufficiently large values of n.

Which topology should be adopted? The question does not arise in finite dimensional contexts. The reason is simply that all 'natural' topologies on finite dimensional Euclidean spaces are *equivalent* in the sense that the corresponding convergence definitions are logically equivalent to one another. This is the case, for example, with the four topologies in the table if they are adapted in the obvious way to a finite horizon context. In all cases, convergence is identical to the usual notion based on the Euclidean metric. Thus the corresponding notions of continuity are also identical.

In contrast, in the infinite horizon model, the noted equivalence fails. It is easily shown that

$$\begin{aligned} d_s - \text{convergence} \quad &\Rightarrow d_\infty - \text{convergence} \\ &\Rightarrow \text{Mackey} - \text{convergence} \\ &\Rightarrow d_p - \text{convergence}. \end{aligned} \tag{2}$$

But none of the reverse implications is true. For example, define the sequences $\{x^n\}$, $\{y^n\}$, and $\{z^n\}$ as follows:

Impatience, Table 1

Topology	Definition of convergence of $\{x^n\}$ to x	Metric
Product	$x_1^n \xrightarrow[n\to\infty]{} x_t$ for all t	$d_p(x,y) = \sup_t \dfrac{2^{-1}\lvert x_t - y_t\rvert}{\{1 + \lvert x_t - y_t\rvert\}}$
Mackey	$\sup_t \lvert a_t \cdot (x_t^n - x_t)\rvert \xrightarrow[n\to\infty]{} 0$ for all sequences of real numbers $\{a_t\}_0^\infty$ that converge to 0	—
Supremum	$\sup_t \lvert x_t^n - x_t\rvert \xrightarrow[n\to\infty]{} 0$	$d_\infty(x,y) = \sup_t \lvert x_t - y_t\rvert$
Svensson	$\sum_0^\infty \lvert x_t^n - x_t\rvert \xrightarrow[n\to\infty]{} 0$	$d_s(x,y) = \min\left(1, \sum_0^\infty \lvert x_t - y_t\rvert\right)$

$$x_t^n = 0 \quad \text{if} \quad 0 \le t \le n \quad \text{and} \quad = n \quad \text{if} \quad t > n$$
$$y_t^n = 0 \quad \text{if} \quad 0 \le t \le n \quad \text{and} \quad = 1 \quad \text{if} \quad t > n$$
$$z^n = (n^{-1}, n^{-t}, n^{-1}, \ldots).$$

Then $\{x^n\}$ converges to (0, 0,...) in the product topology but not in the Mackey topology. In the former case x^n is viewed as being close to the zero consumption path for large n, because the first n generations all have zero consumption. Thus the product topology discounts the fact that in x^n infinitely many generations enjoy large consumption levels which are unbounded as n grows. It is the latter feature which explains why x^n and (0, 0...) are not viewed as being close to one another by the Mackey topology. (Take $a_t = t^{-1/2}$ in the definition of Mackey convergence.) Thus, for example, in the case of $\{y^n\}$ where the consumption of future generations is bounded in n, the sequence is Mackey-convergent to the zero consumption path. The sequence $\{y^n\}$ is not d_∞-convergent since not all generations have consumption near 0. Finally, $\{z^n\}$ converges to (0, 0...) in the sup topology, but it is not d_s-convergent since the 'aggregate' deviation of consumption levels between the two paths is large (indeed $\Sigma_0^\infty |z_t^n| = \infty$).

When topologies are not equivalent continuity of a preference relation has different meaning depending upon which topology is adopted. Thus (2) implies immediately that

$$\begin{aligned} d_p - \text{continuity} \quad &\Rightarrow \text{Mackey} - \text{concergence} \\ &\Rightarrow d_\infty - \text{continuity} \\ &\Rightarrow d_s - \text{continuity}, \end{aligned} \tag{3}$$

and none of the reverse implications is valid. In finite dimensional analysis continuity is a purely technical assumption which is innocuous from an economist's point of view. But the discussion of convergence in the above four topologies strongly suggests that in infinite horizon models the specification of a topology and the assumption of continuity can have economic content. Indeed, continuity in some topologies can imply impatience.

One demonstration of the crucial role played by a topology is provided by Diamond (1965) and Svensson (1980). Call a preference relation *equitable* if it provides equal treatment for all generations in the sense that for all x and y in L_+^∞, $x \succcurlyeq y \Leftrightarrow x \succcurlyeq \pi y$, where πx (or πy) is obtained from x (or y) by permuting finitely many of its components. A preference relation is weakly monotonic if $x_t \succcurlyeq y_t$ for all $t \Rightarrow x \succcurlyeq y$. Diamond shows that there does not exist an equitable and weakly monotonic preference relation that is also continuous in the product metric. This preclusion of equity is perhaps not surprising given the discounting of the future that is built into the definition of d_p. But even given the apparently 'time neutral' metric d_∞, the scope for equity is limited. Diamond proves that equity and d_∞-continuity are incompatible given strong monotonicity ($x_t \ge y_t$ for all t and $x_\tau > y_\tau$ *for some* $\tau \Rightarrow x > y$.) If only weak monotonicity is imposed, then all postulates are satisfied by the maximin ordering, whereby

$$x \succcurlyeq y \Leftrightarrow \inf x_t \ge \inf y_t. \tag{4}$$

The view, based on finite dimensional analysis, that continuity is an innocuous technical assumption, would lead one to interpret Diamond's results as demonstrating the non-existence of equitable orderings that satisfy minimal additional regularity conditions. But, the correct interpretation is the Diamond's theorems demonstrate the strong ethical content of d_p-continuity and d_∞-continuity. The latter view is fortified by Svensson (1980). He shows that if the d_s metric is adopted, then there exist equitable and strongly monotonic orderings which are d_s-continuous. Since d_s is a priori plausible, the onus is clearly shifted to the metric. At the extreme, continuity can be imposed with total impunity if the metric d_0 is adopted, where

$$d_0(x, y) = \begin{bmatrix} 0, & \text{if} & x = y \\ 1, & \text{if} & x \ne y \end{bmatrix}.$$

The topology corresponding to d_0 is called the discrete topology. According to this metric

distinct consumption paths cannot be close to one another, so continuity is automatic. A natural open question is the characterization of metrics d (and more general topologies) such that d-continuous, equitable and (weakly or strongly) monotonic preference relations exist.

At this point it is worth recalling a principal reason that continuity is of interest – namely that by (an extension of) the Weierstrass Theorem, it will guarantee the existence of optimal elements in compact sets. Given a topology Γ on L_+^∞, a set $K \subset L_+^\infty$ is *Γ-compact* if every (generalized) sequence of points in K has a (generalized) subsequence that converges according to Γ to a point in K. As the topology changes in such a way as to permit more continuous functions, the family of compact sets shrinks (see (2) and (3)). Thus as continuity becomes easier to achieve it also becomes less significant. (For example, K is *d_0-compact only if it consists of finitely many points; and there exist many economically relevant sets K* that are compact in the product topology but not in the sup topology. One example arises in an exhaustible resource model where feasible consumptions plans satisfy $\Sigma_0^\infty x_t \leq w$, and w is the initial stock of the good.) If there is a class of constraint sets where the existence of optimal elements is desired, then the 'useful' topologies are those that make each of the constraint sets compact. This approach (emphasized by Campbell, 1985) would remove some of the arbitrariness from the choice of a topology.

Diamond's results suggest that continuity may imply 'some form of impatience', since equity can be viewed as the lack of impatience. A more precise definition of impatience is required for a clearer demonstration of the link between the latter and continuity. For example, impatience could be taken to mean that interchanging the consumption levels of generations 1 and t results in a strictly preferred plan if period t consumption was initially larger. If the preceding statement is valid only for t sufficiently far into the future, then *eventual impatience* could be said to prevail. This latter notion captures not only a preference for the advancement of the timing of satisfaction, but also the idea that the taste for future consumption diminishes as the time of consumption recedes into the future. These and related definitions appear in Koopmans (1960), Koopmans et al. (1964) and Diamond (1965). Their proofs that appropriate continuity implies (eventual) impatience depend, with the single exception of Diamond (p. 174), on maintained separability assumptions on the preference relation. The separability assumptions can be deleted if the existence of a differentiable utility function is assumed (Burness 1973).

Brown and Lewis (1981) define some notions of asymptotic impatience. For example, they call a preference relation *strongly myopic* if for all x, y and z in L_+^∞, $x \succ y \Rightarrow x \succ y + {}_nz$ for all sufficiently large n, where ${}_nz = (0,\ldots,0,z_{n+1}, z_{n+2}, \ldots)$. In other words, the preference for x over y is unchanged by an increase in the latter programme in the consumption of infinitely many generations, as long as the increase occurs only for generations that are situated sufficiently far into the future.

Interpret a preference relation as belonging to a consumer rather than to a central planner. Consumption programmes in L_+^∞ descendants; the latters' consumption levels matter because of intergenerational altruism. This is a common framework in the capital theory literature where the behavioural assumption of impatience is often maintained. This suggests that from the perspective of capital theory, economically interesting topologies are those which (through continuity) imply myopia. For example, any preference relation which is d_p-continuous is necessarily strongly myopic. But the implication is false if the product metric is replaced by d_∞ or d_s. Brown and Lewis show that the Mackey topology bears a special relationship to strong myopia. Mackey-continuity is the weakest continuity requirement (corresponding to topologies in a broad and convenient class) that can be imposed on a preference relation in order that strong myopia be implied. Thus it is a 'natural' topology if strong myopia is desirable.

There is an important link between the Mackey topology and strong myopia on the one hand and general equilibrium analysis in the framework of

'infinitely lived' agents on the other. Bewley (1972) points out that the Mackey topology is particularly appropriate for general equilibrium analysis because continuity requirements weaker than Mackey-continuity do not guarantee the existence of equilibria with price systems that can be represented by absolutely summable sequences $(p_0, \ldots, p_t, \ldots)$, rather than merely for more general mathematical constructs that have no economic interpretation. In light of the relationship between Mackey-continuity and strong myopia, the latter seems necessary for meaningful general equilibrium analysis.

Brown and Lewis sharpen the link between impatience and general equilibrium analysis. They prove that if individual preferences are suitably monotonic, then Mackey-continuity and strong myopia are unnecessarily strong assumptions. But a form of asymptotic impatience is still relevant. Call a preference ordering *weakly myopic* if the implication defining strong myopia is valid for all constant programmes z. Then even if individual preferences are weakly monotonic, the existence of economically interpretable equilibrium price systems as above can be guaranteed only by continuity requirements which imply weak myopia.

Suppose that we are willing to accept more general constructs (linear functionals on L_+^∞) as price systems. Can we then dispense with impatience? Araujo (1985) provides a negative partial answer. He restricts attention to a well-defined subset of those continuity conditions which lie 'between' d_∞-continuity and d_p-continuity. Then he shows that the existence of such general price systems can be guaranteed only if continuity requirements are imposed which imply strong myopia, or, when suitable monotonicity is maintained for preferences, weak myopia. Existence of equilibria cannot be guaranteed in such cases as the maximin ordering (4) which exhibits no impatience.

We offer one final comment. In a planning context, continuity of the social preference relation may be desirable not necessarily for its own sake nor because it may imply myopia, but primarily to guarantee that the preference relation be *effective*, that is, that optimal consumption paths exist. From this perspective, it seems more pertinent to investigate the link between effectiveness and impatience directly, without involving continuity which is, after all, at best sufficient and definitely not necessary for the existence of optimal paths. Thus, for example, a pertinent question is whether impatience (in some precise sense) is necessary for effectiveness in a relevant set of choice environments. While this question has been addressed to some extent in the growth theory literature cited earlier based on the additive utility functional (1), an analysis comparable in generality to that of Brown and Lewis or Araujo has yet to be performed.

See Also

- ▶ Fisher, Irving (1867–1947)
- ▶ Present Value
- ▶ Time Preference

Bibliography

Araujo, A. 1985. Lack of Pareto optimal allocations in economies with infinitely many commodities: The need for impatience. *Econometrica* 53(2): 455–461.

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4(3): 514–540.

Brown, D.J., and L.M. Lewis. 1981. Myopic economic agents. *Econometrica* 49(2): 359–368.

Burness, H.S. 1973. Impatience and the preference for advancement in the timing of satisfactions. *Journal of Economic Theory* 6(5): 495–507.

Diamond, P.A. 1965. The evaluation of infinite utility streams. *Econometrica* 33: 170–177.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Koopmans, T.C. 1960. Stationary ordinal utility and impatience. *Econometrica* 28: 287–309.

Koopmans, T.C. 1966. On the concept of optimal economic growth. In *The econometric approach to development planning*. Amsterdam: North-Holland.

Koopmans, T.C. 1967. Objectives, constraints, and outcomes in optimal growth models. *Econometrica* 35: 1–15.

Koopmans, T.C., P.A. Diamond, and R.E. Williamson. 1964. Stationary utility and time perspective. *Econometrica* 32: 82–100.

Magill, M.J.P. 1981. Infinite horizon programs. *Econometrica* 49(3): 679–711.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.
Svensson, L.G. 1980. Equity among generations. *Econometrica* 48(5): 1251–1256.
von Böhm-Bawerk, E. 1912. *Positive theory of capital*. South Holland: Libertarian Press, 1959.
von Weizsäcker, C.C. 1965. Existence of optimal programs of accumulation for an infinite time horizon. *Review of Economic Studies* 32: 85–104.

Imperfect Competition

Louis Makowski

Imperfect competitors are individuals or firms who face downward-sloping demand curves or upward-sloping supply curves for some product (s). This is to be contrasted with perfect competitors who, by definition, face perfectly elastic demand and supply curves for all products. Notice we define perfect competitors not just as price-takers, but as rational price-takers: perfect competitors cannot influence the levels of market clearing prices. By contrast imperfect competitors, by their presence, can influence some equilibrium prices. As simple as these definitions sound, they hold within themselves a world of meaning that we will explore a little in this entry.

Since the early days of economics as a science, the importance of the force of competition has been stressed. Adam Smith viewed the force of competition as a central benefactor of society, which both (a) guards people against the possibility of monopolistic exploitation by insuring that the long run price will not exceed the cost of production; and (b) automatically provides for long-run progress by firing entrepreneurs' restless search for new profit potentials. In contrast to Smith, modern-day economists are becoming increasingly uncertain whether the force of competition is entirely beneficent. The image of wasteful competition between individuals and between firms is gaining repute. Theories of imperfect competition are becoming increasingly popular, reflecting a dissatisfaction with the predictive power of the perfectly competitive model of economic reality.

The insight that competition can be wasteful, not necessarily beneficent, was popularized by Edward Chamberlin, who along with Joan Robinson is typically credited with renewing economists' interest in imperfect competition beginning in the 1930s (Chamberlin 1933; Robinson 1933). As a contender to the perfectly competitive image of economic reality, Chamberlin offered his image of many firms selling differentiated products, contending with one another, but nevertheless each facing a downward-sloping demand curve. His famous 'excess capacity theorem' was the caricature he offered of wasteful competition.

As in Chamberlin, the current modelling of imperfect competition tends to be partial equilibrium. A popular practice is to make the assumption that firms will interact in a Cournot–Nash fashion. Perhaps more ambitious and interesting are current explorations at the interface of game theory and industrial organization theory. Many (small group) models of imperfectly competitive interactions are available, each with its own idiosyncratic, stylized features. These models are a beginning toward analysing imperfect competition between individuals and between firms as an active process. But there does not currently exist a standard paradigm of imperfect competition (either partial equilibrium or general equilibrium). This contrasts sharply with the case of perfect competition, which is typically idealized using a Walrasian general equilibrium model. Perhaps models of imperfect competition must necessarily be legion and case-specific?

We will not try to survey existent models of imperfect competition here. Rather, we will try to offer some overview in terms of a unifying principle. In particular, we will argue that *increasing returns* is the usual source of imperfect competition. Knowledge of such a unifying source will hopefully help the reader make sense of the plethora of available idiosyncratic models. It should also help the reader understand why imperfect competition, in contrast to perfect competition, may be wasteful.

The Meaning of Increasing Returns: From Perfect Competition to Imperfect Competition

To understand the concept of increasing returns, as applied to the economy as a whole (rather than to a particular firm), it is useful to first understand how economists usually ensure that a model of the economy as a whole will be *perfectly competitive*. This will provide us with a benchmark from which to proceed since, as we shall see, a perfectly competitive economy typically exhibits constant returns, in contrast to increasing returns. (The observation that constant returns typifies perfect competition is also central to Samuelson (1967). He proceeds in a somewhat different fashion, but his article may be read as a useful complement to this one.) To ensure that an economy will be perfectly competitive economists typically assume a finite number of homogeneous private goods. Then, keeping the set of goods fixed, they *replicate* the economy by increasing the number of buyers and sellers of each commodity indefinitely. The resulting, limiting economy will be perfectly competitive in the sense that the force of competition between the many alternative sellers of each commodity and the many alternative buyers will be sufficient to ensure that no one individual will possess any monopoly or monopsony power. That is, no one individual will be able to influence the levels of the prices that equilibrate supply and demand. For example, if some seller tries to exploit some buyer there will be plenty of perfect substitute sellers available ready to take the buyer away from him. Notice that the image of 'thick markets', i.e. homogeneous private goods with many small sellers and buyers of each good, is central to economists' image of perfect competition. It is this image of thick markets that Chamberlin found to be a grotesque caricature of our economic reality. It is easy to see that a large, replicated private-good economy exhibits constant returns to scale in the sense that a small subset of its participants could do as well on their own as they could participating in the economy as a whole. The economy can be 'disintegrated' without loss of consumers' surplus or gains from trade.

The analogy to ordinary production theory can be made more precise in an idealized special case, that of transferable utility – where utility can be regarded as cardinal and additive over individuals. (Notice this is essentially equivalent to assuming that everyone always enjoys constant marginal utility from income.) In this case, one can construct an analogy to an ordinary firm production function for the economy as a whole (a sort of 'aggregate production function'), and one can show that in the limit, replication will result in this function exhibiting constant returns. Further, in a perfectly competitive equilibrium all individuals will be rewarded with their marginal products to the economy as a whole, calculated from this 'aggregate production function'.

This idealized special case is useful for gaining parable-like insights into the nature of not only perfect competition, but also imperfect competition. So we shall first sketch some of the claims made for it above (for further details, see Makowski and Ostroy 1987). The basis for its usefulness is that, if we assume utility is cardinal and additive over individuals, then we can formalize the idea that the economy as a whole is in the business of producing *utility* for its participants. In particular, with this assumption we can let g(S) equal the total potential gains from trade possible in a subeconomy consisting only of the set of individuals S; i.e. g(S) equals the maximum total utility achievable by S when it can only trade within itself. Then we can regard g, the total potential gains from trade *function* (defined over all possible subeconomies S) as the economy's 'aggregate production function'. Notice that the range of g is defined in utility space: the economy as a whole produces utility as its output. And the domain of g is subsets of individuals: individuals are the 'inputs' used to produce utility, by exploiting the gains from trade. (In cooperative game theory, the g function would be called a 'characteristic function'. But we shall restrict our attention to non-cooperative, bilateral interactions; this may be rationalized by assuming that multilateral coalition formation is prohibitively costly.)

Just as with any production function, we can define the marginal product of each factor of

production – now each individual rather than each commodity since the domain of g is subsets of individuals. In particular, it is natural to define the potential marginal product of individual i to the economy as a whole, MP_i, as the difference between the potential gains from trade in the economy as a whole, $g(A)$ (where A is the set of all individuals) and the potential gains from trade in the absence of individual i, $g(A^i)$ (where A^i is the set of all individuals in the economy except i); i.e. $MP_i \equiv g(A) - g(A^i)$. Notice MP_i just equals individual i's contribution to the total potential gains from trade in the economy.

It can be shown that in any perfectly competitive economy, each individual's final utility level (say u_i) just equals his potential marginal product to the economy as a whole. That is, the total gains from trade are distributed under perfect competition that $u_i = MP_i$ for each individual i. Thus the analogy to ordinary production theory under perfect competition, where each factor earns its *MP*, is complete. Since any perfectly competitive equilibrium is efficient (i.e. the actual gains from trade equal the maximum potential gains), this implies there must be 'adding-up' in any perfectly competitive economy: the sum of all individuals' *MP*'s to the economy as a whole must equal the total potential product of the economy, g(A).

Constant returns and adding-up are intimately related. Both are achieved by replication as follows. Typically the above g function will initially exhibit increasing returns in the sense that the sum of all individuals' *MP*'s will exceed the total potential 'output'. But, for larger and large economies this sum approaches $g(A)$. The process is idealized in the limit – when we can regard individuals as infinitesimal, i.e. points on a line. In this limiting, continuum-of-individuals case the g function will be homogeneous: multiplying all 'inputs' by any factor will just multiply the total achievable gains from trade by the same factor. Hence, 'adding-up' in the limit is ensured by Euler's Theorem. (Individual i's potential marginal product in the limiting, continuum economy just equals the partial derivative of g with respect to that individual, evaluated at A, rather than the finite different $g(A) - g(A^i)$.)

Thus, the connection between replication, constant returns, and the nature of perfectly competitive economies is clarified. In particular, we now see that such economies exhibit, in the limit, constant returns *over* (the 'inputs') *individuals*. One deeper result from perfect competition theory will also be useful, before we leave this benchmark case for the domain of imperfect competition. It can be shown that not only does perfect competition imply

(i) $u_i = MP_i$ for each individual i; and (ii) $\Sigma MP_i = g(A)$, but conversely, (i) and (ii) also imply perfect competition. Thus, perfectly competitive economies are essentially *equivalent* to ones in which constant returns over individuals prevails. In the absence of such constant returns, we could not rely on Euler's Theorem to ensure adding-up, (ii); consequently, it would be a mere accident if one could reward everyone with their *MP*'s to the economy as a whole.

This last, equivalence observation provides us with a key for transiting into the realm of imperfect competition. Since the presence of constant returns over individuals essentially characterizes perfectly competitive economies, its absence essentially characterizes economies without perfect competition, i.e. economies in which competition must necessarily be *imperfect*. But under what circumstances will competition necessarily be imperfect? Or, expressed in terms of our idealized special case, under what circumstances will the g function not exhibit constant returns over individuals?

The replication image of perfect competition gives us our first insight into such imperfectly competitive economies. They are economies in which there are not sufficient perfect substitute sellers or buyers for the force of competition to ensure that no individual can influence the levels of market clearing prices. But what does this mean in terms of our gains from trade function? As noted above, in the absence of perfect competition (e.g., in small economies) the g function will typically exhibit *increasing returns* over individuals, in the sense that the sum of all individuals' *MP*'s will typically exceed the total potential gains from trade. To illustrate with a paradigmatic example of imperfect competition – bilateral

monopoly – consider an economy with just one buyer and one seller, and with potential gains from trade between them. Then each individual is *crucial* to realizing the gains from trade. In particular, without either there would be zero gains from trade, so the *MP* of *each* equals the total potential gains from trade, $g(A)$. But then the sum of their *MP*'s exceeds the total potential gains from trade since $\Sigma MP_i = g(A) + g(A) = 2\,g(A)$. So, there are increasing returns over individuals in bilateral monopoly situations. Obviously each person cannot appropriate all of $g(A)$.

That the sum of the two individuals' *MP*'s exceeds the potential gains from trade between them has the following interpretive significance. Imagine the buyer and seller contending with one another over their respective shares of the total economic pie, $g(A)$. Each might insist on receiving his full potential contribution to the size of the pie, his *MP*. But in cases of imperfect competition, this is impossible to achieve. (Note that, by contrast, under perfect competition each seller (respectively, buyer) receiving his full *MP* would be the *inevitable outcome of competition* between alternative competing buyers of the seller's output (respectively, alternative competing sellers to the buyer). The consequence in terms of prices is that under perfect competition no one buyer (respectively, seller) can influence the level of market clearing prices.) We might next imagine each individual engaging in devious bargaining tactics to win at least as much of the pie for himself as he can. Such manoeuvrings are generally resource costly, hence the whole size of the pie may well diminish in the process of bargaining for shares of it. This is the image of wasteful competition! Our story indicates how increasing returns over individuals, and the consequent failure of adding-up of individuals' *MP*'s to the economy as a whole, can give rise to wasteful competition. That the potential economic pie cannot be naturally imputed to individuals, via their contributions to the size of the pie (their *MP*'s), makes the potential gains from trade a common property resource to be contended over wastefully.

An Example of Wasteful Competition

To make the discussion more concrete, we now present a more explicit example involving bilateral monopoly. Imagine an economy with just one barber B and one customer, C. B can cut hair costlessly, and C is willing to pay up to w dollars for one haircut (he does not want more than one); hence $g(A) = w$ which, recall, also equals each individual's *MP*. Will the full potential gains from trade be realized?

Suppose at the beginning of the world nature picks C's willingness to pay for a haircut from a distribution between 0 and 10, so that any w in this interval is an equally likely choice by nature. Suppose further that Mr C knows his actual type, w, but Mr B only knows the distribution from which nature has picked C's type. Then bilateral bargaining will not generally result in all the potential gains from trade being realized. To see why suppose B is a tough bargainer and can commit himself to a take-it-or-leave-it price for a haircut.

Then, given his incomplete information about C's type, it is easy to see he will commit himself to a price of \$5/haircut; this maximizes his expected profits. But then, whenever C's true willingness to pay is less than \$5, he will not get a haircut although it is efficient for him to do so given B's cost of haircuts is zero; $g(A)$ will not be realized. For example, suppose $w =$ \$4, then although C may go to B and say 'I am willing to get a haircut if you will lower the price to something less than \$4,' B will rationally not believe him and change his price, since if he believed C in this case then C would rationally pretend to have a w less than \$5 even when his true w is greater than \$5.

Notice that the basic source of the inefficiency when $w <$ \$5 is the potential deviousness by C about his true willingness-to-pay – in an effort to induce a lower price and hence a bigger share of his full potential marginal product, w – coupled with B's contrary effort to extract the biggest possible share of *his* potential marginal product, w, by making a price commitment that reflects his ignorance about w. Summarizing, (wasteful)

competition between B and C over the potential gains from trade results in the actual gains, zero, falling short of the potential gains, *w*, whenever *w* $5. Wasteful competition is reflected in the *underproduction* of haircuts.

In contrast, notice that in a replicated economy with many identical B's and C's, (perfect) competition between barbers for customers would force the price of haircuts down to their true cost, zero. Hence, the full potential gains from trade would be realized without devious, wasteful competition. (The reader can check that in this replicated case the *MP* of any one barber equals zero while that of any one customer equals *w*; hence there is 'adding-up' in this case.)

The fact that imperfect competition is generally inefficient – it frustrates Adam Smith's Invisible Hand – is so central to our understanding of the economic import of imperfect competition that it is perhaps useful to re-phrase the source of market failure under imperfect competition in terms of 'externalities' since it is well-understood that externalities give rise to market failures. Under perfect competition each individual appropriates his full potential contribution to society, his *MP*. Consequently, he creates *no externalities*, beneficial or harmful, to others. By contrast, under imperfect competition not everyone can appropriate his full potential contribution to society, his MP. Consequently, if an equilibrium allocation with imperfect competition is to be efficient, some individual(s) must create external benefits *for others* (since some individuals must receive less than their marginal products). But no one cares about *external* benefits, only about the benefits he can internalize (i.e. appropriate). Consequently, in trying to internalize as much of his contribution as possible, an imperfect competitor will engage in wasteful market tactics most of whose harmful consequences others must bear.

Multilateral examples of imperfect competition, more in the spirit of Chamberlin, can also be constructed. In such examples, increasing returns lead to the gains from trade between producers and consumers being a common property resource that cannot be naturally imputed to agents using the *MP* reward principle. This can lead to 'excess capacity' as some industries' potential profits become a common property resource to be contended over wastefully via over-entry. In contrast, under perfect competition entry is efficiently guided since each firm's profits just reflect *its MP*; not any share of some other firm's potential *MP* that it can steal away by entering the industry.

Notice that throughout this article we are supposing there do not exist any non-market external effects between economic agents. So, all interactions are voluntary and involve exchange. But this does not exclude the possibility of external effects between economic agents in their trade relationships, so called 'pecuniary externalities'. Indeed, the possibility of such trade-related externalities is the essence of imperfectly competitive interactions and the source of the Invisible Hand's failure to achieve Pareto efficient outcomes under imperfect competition. (A terminological note: We refer to imperfect competition as 'wasteful' relative to the benchmark of achieving pure Pareto efficiency. A related question that we do not address in this entry is: Can one find institutions that could improve on the market outcome in the presence of imperfect competition? Some economists would argue that the answer is 'no'; hence that the market outcome provides the best *realistic* benchmark even in the presence of imperfect competition, for example see Demsetz 1959.)

Indivisibilities, Complementarities and Increasing Returns

There is a tradition in economic theory that views some sort of indivisibility as the main source of increasing returns. In this tradition, if just doubling the amounts of all factors results in more than double the output, the source of increasing returns is interpreted in terms of indivisibilities in some specialized functions of factors.

That indivisibilities are the usual source of increasing returns was disputed by Chamberlin in a famous controversy with Kaldor; the latter

subsequently recanted his position (see Kaldor 1972). Without clouding ourselves in the smoke raised by this issue, we can shed some light on the central substantive aspect. At the heart of the dispute is the question, will sufficiently large economies necessarily be perfectly competitive? (Notice that the idea of indivisibilities suggests that at some sufficiently large level of production all scale economies will be exhausted.) Thus it is interesting to observe that increasing returns can exist even in large economies.

In particular, *how* one replicates an economy is crucial to whether a replicated economy will become closer and closer to a perfectly competitive one. For perfect competition to result in the limit, (1) it is essential that one only allows private goods, not collective goods: replicating an economy with collective goods generally does not diminish the presence of monopsony power on the buyers' side since each buyer never competes with other buyers for units of a *collective* good. This monopsony power gives rise to manifestations of wasteful competition by each buyer – to try to appropriate the biggest possible share of his contribution to the gains from trade, his *MP* – such as 'free rider problems'.

(A bibliographical note: Samuelson introduced the concept of collective goods to Anglo-American economists in a series of articles (Samuelson 1954, 1955 and 1958). He forcefully argues that public goods differ fundamentally from private goods insofar as the ability of the Invisible Hand to allocate them efficiently is concerned. One can detect, in reading his three articles chronologically, a maturing in Samuelson's appreciation of the source of market failure as being due to some sort of increasing returns in public good economies. This point is made in Head (1962), whose article may be read as a useful complement to this one. Head stresses difficulties in appropriation as the source of market failure with collective goods, without explicitly using the MP concept.)

More in the spirit of Chamberlin, (2) it is also essential to keep the set of private commodities relatively fixed while one replicates: if the set of commodities expands at the same rate as the set of buyers and sellers, then perfect competition need not emerge even in the limit. Some sellers may still be 'special' as far as some buyers are concerned; thus a seller may still face a downward sloping demand curve reflecting the tastes of buyers who regard the seller's product as special (e.g., see Hart 1985). In this context, the right image of a large economy is an ever-expanding nexus of complementarities between individuals, that never becomes large enough to be 'disintegrated' without loss in potential gains from trade (Kaldor 1972, and Allyn Young's classic 1928 paper may be usefully read on this point). In this image the possibilities for increasing returns are never exhausted since essential complementarities between individuals are never exhausted. Notice that the reason for increasing returns here is more easily explained in terms of the existence of *complementarities* between individuals, rather than indivisibilities. Expressed in terms of our idealized special case, as long as there exist essential complementarities between individuals, the gains from trade function will continue to exhibit increasing returns *over individuals*. In this common case, the force of competition will not be sufficient to guarantee that everyone has perfect substitutes. Thus competition between individuals may remain imperfect – and wasteful – even in large economies.

See Also

- ▶ Competition
- ▶ Entry and Market Structure
- ▶ Monopolistic Competition
- ▶ Perfectly and Imperfectly Competitive Markets

Bibliography

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Demsetz, H. 1959. The nature of equilibrium in monopolistic competition. *Journal of Political Economy* 67: 21–30.

Hart, O.D. 1985. Monopolistic competition in the spirit of Chamberlin. *Review of Economic Studies* 52: 529–546.

Head, J.G. 1962. Public goods and public policy. *Public Finance/Finances Publiques* 17(3), 197–219.

Kaldor, N. 1972. The irrelevance of equilibrium economics. *Economic Journal* 82(December): 1237–1255.
Makowski, L., and J.M. Ostroy. 1987. *Vickrey–Clarke-Groves mechanisms and perfect competition*. June: Journal of Economic Theory.
Robinson, J. 1933. *Economics of imperfect competition*. London: Macmillan.
Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36-(November): 387–389.
Samuelson, P.A. 1955. Diagrammatic exposition of a theory of public expenditure. *Review of Economics and Statistics* 37(November): 350–356.
Samuelson, P.A. 1958. Aspects of public expenditure theories. *Review of Economics and Statistics* 40-(November): 332–338.
Samuelson, P.A. 1967. The monopolistic competition revolution. In *Monopolistic competition theory: Studies in impact. Essays in Honor of Edward H. Chamberlin*, ed. R.E. Kuenne, New York: Wiley.
Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38(December): 527–542.

Imperfectionist Models

John Eatwell

The term 'imperfectionist' was applied by Eatwell and Milgate (1983) to those models which rely on imperfections or arbitrary constraints in order to analyse the phenomenon under consideration. In other words, an imperfectionist analysis involves the construction of a model which, when innocent of those arbitrary constraints, does not display the phenomenon. The leading species of this genus to be found in economics today are models of unemployment in which imperfections such as sticky prices, or the effects of uncertainty, are imposed on a Walrasian model, thus disrupting the Walrasian relationship between price formation and the determination of levels of output which implies clearing of the markets for endowments of factor services.

The key issues in any consideration of the relationship between the theory of output and the theory of value and distribution can be revealed by the answers given to two questions:

(1) Does the determination of relative prices in a market economy also involve the determination of the size and composition of output and, in particular, is the level of output such that labour is fully employed (in the sense that at the going wage all workers willing to offer labour would be able to find employment)?
(2) Are variations in relative prices associated with variations in output such that the economy tends towards a level of output compatible with the full employment of labour?

Each of these questions can be supplemented with a further question: if not, why not?

The significance of these questions can be illustrated in terms of the most elementary piece of orthodox neoclassical analysis. According to this account, 'equilibrium' is determined at the point of intersection of a function relating price to quantity demanded and another relating price to quantity supplied. When this view of price determination is extended to the economic system as a whole, the equilibrium position of the economy is characterized by a set of market-clearing prices, with associated quantities (levels of commodity output and levels of 'factor' utilization), such that the markets for all commodities and all 'factors of production' clear. In particular, the labour market clears at the equilibrium level of the wage (relative to the associated set of equilibrium prices).

In terms of this familiar approach to the analysis of price formation the answer to the first question is obvious. Equilibrium prices and equilibrium quantities are determined simultaneously. The theory of value, based on demand and supply, is one and the same thing as the theory of output. If there exists an equilibrium set of prices then there exists an equilibrium set of outputs – equilibrium in the sense of market clearing, including the full employment of labour, as defined above. Furthermore, this theory of the simultaneous determination of prices and quantities is typically presented in such a way – by juxtaposing demand and supply functions – that the idea that prices adjust automatically so as to clear markets, thus tending to push the economic system towards a full-employment level of output, seems to follow as a self-evident corollary of the

theory. (It does not in fact follow as readily as might appear at first sight, since the stability of an equilibrium is far more difficult to demonstrate than its existence.)

Here, then, one has the demand-and-supply (neoclassical) analysis of prices and quantities in a nutshell: the equilibrium set of outputs (and levels of 'factor' utilization) is determined simultaneously with the equilibrium set of prices (of commodities and 'factors of production'); variations in relative prices sparked off by an imbalance between demand and supply, will be associated with variations in quantities in a direction which ensures that both prices and quantities tend towards their equilibrium levels. Neoclassical analysis, therefore, answers the first two questions posed above in the affirmative.

An analysis of unemployment may then be derived directly from these relationships between prices and quantities. Any *inhibition* to the tendency of prices and quantities to find their equilibrium (market-clearing) levels will leave the economic system in disequilibrium with, perhaps, either an excess demand for labour or an excess supply of labour (ie unemployment). An enormous variety of analyses of unemployment are constructed in this way.

The general tenor of the neoclassical analysis of the causes of unemployment is that while the economy would be self-regulating in the best of all possible worlds (ie the implicit tendency towards the full employment of labour would be realized) – the market is *inhibited* from fulfilling this task by the presence of certain 'frictions' or 'rigidities'. In the literature on the problem of unemployment, examples of such inhibitions are legion. They include: 'sticky' prices (particularly 'sticky' or even rigidly fixed wages and/or 'sticky' interest rates); institutional barriers to the efficacy of the price mechanism, such as monopoly pricing (by firms or individual groups of workers); inefficiencies introduced into the working of the 'real' economy by the operations of the monetary system; the failure of individual agents to respond appropriately to price signals because of disbelief in those signals, the disbelief being derived from uncertainty about the current or future state of the market, or from incorrect expectations concerning future movements in relative prices, or from false 'conjectures' about the actual state of the market.

Indeed, examples of 'frictions' and 'rigidities' can be multiplied at will – any factor which causes the market to work *imperfectly* will do. It will be convenient, therefore, to group all the authors of the myriad of arguments of this kind together under the general heading of 'imperfectionists'. (It should be noted that by referring to this kind of analysis as 'imperfectionist' I do not intend to imply that the envisaged failure of the market mechanism to operate in the way depicted by the underlying demand-and-supply theory *necessarily* derives from imperfections of competition.)

Underlying them all is a fundamental similarity: that if the particular aspect (or aspects) of the economic system which gives rise to the breakdown of the market mechanism were to be absent, then the system would tend towards the full employment of labour (and other 'factors of production'). Thus, in all cases, the analysis of unemployment is viewed as no more than an aspect of the neoclassical theory of value and distribution. According to this approach, whether a relatively 'optimistic' or 'pessimistic' stance is taken with respect to the efficacy of the market mechanism in promoting full employment, the analysis of output and employment is part and parcel of the theory of relative price determination. This is so even in the case of those imperfectionists who feel that the essential workings of the theory are distorted gravely in the real world.

In marked contrast to the analysis outlined above are those theories of employment which propose no particular functional relationship between prices and quantities. The central proposition of neoclassical analysis, that the theory of value and distribution is also the theory of output, is rejected, together with the connected notion that appropriate variations in relative prices will promote variations in quantities, so moving the economic system in the direction of a full-employment equilibrium.

Unfortunately, this rejection of the neoclassical theory of value and distribution – of the entire apparatus of demand-and-supply analysis – has not always been backed up by rigorous analytical

argument; so much so that it has sometimes been confused with an imperfectionist position. A striking example of this is the rejection by a number of writers of the neoclassical theory of value, and their advocacy of the idea that relative prices, far from being determined by demand and supply, are determined by a mark-up over normal prime cost where this mark-up is insensitive to variations in the conditions of demand (see, for example, Kalecki 1939; Neild 1963; Godley and Nordhaus 1972). Quite apart from the obvious shortcomings of 'mark-up' analysis as a theory of price formation – it is in essence a proposition about the stability of the ratio between prices and costs rather than a theory about the determination of either of those magnitudes, or even of the size of the ratio – this attempt to separate the study of relative price determination from the analysis of output may readily be confused with an imperfectionist argument based on 'sticky' prices arising from the presence of monopolistic or oligopolistic influences in commodity markets. (Thus Malinvaud (1977) cites the results of Godley and Nordhaus (1972) in support of his orthodox imperfectionist position.) Moreover, the bald assertion that prices and quantities do not bear the well-defined functional relationship to one another that is postulated in neoclassical theory does not provide a satisfactory analytical basis upon which to build up a critique of the neoclassical position.

Yet the requisite critique does exist, and is to be found in the outcome of the debate over the neoclassical theory of distribution and, in particular, over its treatment of 'capital' as a 'factor of production' on a par, so to speak, with land and labour. While this debate is seen by many as a rather esoteric controversy in the more abstract realms of economic theory, its implications are more far-reaching than has hitherto been appreciated. The central conclusion of the debate may be summed up, in broad terms, as follows: when applied to the analysis of a capitalistic economy (that is, an economic system where some of the means of production are reproducible), the neoclassical theory is logically incapable of determining the long-run equilibrium of the economy and the associated general rate of profit whenever capital consists of more than one reproducible commodity. Since, in equilibrium, relative prices may be expressed as functions of the general rate of profit, the neoclassical proposition that equilibrium prices are determined by demand and supply (or, more generally, by the competitive resolution of individual utility maximization subject to constraint) is also deprived of its logical foundation.

The relevance of this critique of the neoclassical theory of value and distribution to the problem of the missing critique of the neoclassical theory of output and employment should be apparent from what has already been said. Because the neoclassical analysis of the determination of prices and the determination of quantities is one and the same theory (that of the mutual interaction of demand and supply), the critique of the neoclassical theory of value is simultaneously a critique of the neoclassical theory of output and employment. Therefore, the first of the two questions that were posed at the very outset of this discussion must, on the grounds of the requirement of logical consistency alone, be answered in the negative. The second question, from which neoclassical theory derives the idea that under the operation of the market mechanism there is a long-run tendency towards a determinate full-employment equilibrium, is rendered superfluous.

But this is not all. If the general (or long-run) case of the neoclassical model has been shown to be logically deficient, then all imperfectionist arguments of the introduction of particular (or short-run) modifications into the general case – are incapable of providing a satisfactory analysis of the problem of unemployment. This is not to say that many of the features of the economic system cited by the imperfectionists will have no role to play in a theory of employment based on quite different foundations to those adopted by the neoclassicals. After all, much of the credibility of imperfectionist arguments derives from their pragmatic objections to the direct applicability of the assumptions of the more abstract versions of demand-and-supply theory. But pragmatism is not enough. The implications of more realistic hypotheses must be

explored in the context of a general theoretical framework within which they are integral parts, not imperfections.

See Also

► Keynesianism

Bibliography

Eatwell, J., and M. Milgate (eds.). 1983. *Keynes's economics and the theory of value and distribution*. London: Duckworth.
Godley, W.A.H., and W.D. Nordhaus. 1972. Pricing in the trade cycle. *Economic Journal* 82: 853–882.
Kalecki, M. 1939. *Essays in the theory of economic fluctuations*. London: Allen & Unwin.
Malinvaud, E. 1977. *The theory of unemployment reconsidered*. Oxford: Basil Blackwell.
Neild, R.R. 1963. *Pricing and employment in the trade cycle*. Cambridge: Cambridge University Press.

Imperialism

Alice H. Amsden

Few subjects of such conspicuous historical importance have so consistently escaped lucid theoretical exposition as imperialism. The neoclassical economists have made no theoretical gains whatsoever in the field, having chosen to ignore the subject altogether. Their starting and ending point is a short essay borrowed from Schumpeter in which imperialism in the nineteenth and twentieth centuries is attributed to the atavism of states, acting on feudal and absolutist impulses from an earlier precapitalist era. The field, therefore, has been dominated by Marxists. 'To write about theories of imperialism is already to have a theory,' states Barratt Brown (1972). In modern times, just to use the word is to label what is said as Marxist. The word – like capitalism itself – also implies a theory of broadly construed economic systems and long historical epochs. The sweep of the subject matter is reflected in the breadth of the two major propositions that Marxists have posed: that imperialism and monopoly capitalism are synonymous; and that capitalism underdevelops the third world. The sweep of the subject matter has lent itself to meaningless generalizations and reductionist arguments. But to ignore imperialism altogether on the ground that it is a political phenomenon is to abrogate a responsibility to study a major dimension of economic life, in particular the relationship between the operations of the market and coercive mechanisms.

Part of the problem lies in the ambiguity of the term. Since there is no agreement on the referent of imperialism, there is none on the meaning of the word itself. Marx and Engels did not discuss imperialism as such so they bequeathed no definition. To one of their followers, Rosa Luxemburg (1913), it was the political expression of the accumulation of capital in its competitive struggle for what is still left of the non-capitalist regions of the world. To another, Nikolai Bukharin (1914), it was a policy of conquest by finance capital that is characteristic of one stage of capitalist development. To a follower of a later generation, Samir Amin (1976), it was the perpetuation and expansion of capitalist relations abroad by force or without the willing consent of the affected people. Schumpeter (1919) defined it as the objectless disposition on the part of a state to unlimited forcible expansion.

While no consensus exists, most definitions share an idea that interactions between two social formations are in some sense imperialist if they depend upon force. And the use of force is all the more likely if the two entities are of unequal strength. This is not to say that only military domination qualifies as imperialism. Or that any exchange, commercial or financial, between two parties of unequal strength is imperialism. Rather, even if the use of force is only implicit, perpetrated by the fountain pen, it qualifies as imperialist if the weaker collectivity is subjected to some sort of control by the stronger. So defined, and such is the definition followed below, imperialism is ultimately a political phenomenon, whatever its underlying tap-root.

There appear to be as many explanations for the motivations underlying imperialism as there have been wars. Yet the economic explanations are qualitatively distinct from the rest – geopolitical, psychological – because they reflect the fact that different economic systems reproduce themselves differently. In societies where reproduction was constrained by the availability of land, territorial expansion was the impetus. In societies dependent upon slavery, there was warring for slaves. To buy cheap and sell dear in the age of mercantilism, there was resort to plunder. Come the capitalist system, imperialism evolved into something more complex than theft. It was embodied in exchange relationships. And since exchange could occur peacefully, without the use of force, some, like Schumpeter, presumed that capitalism and imperialism were antithetical. Yet force has been used to accelerate the onset of exchange relationships, to preserve them, and to improve the terms of exchange. Imperialism under centralized planning involves still another dynamic, since the driving imperative for markets (for economic surplus) is absent. It has been attributed by Ota Sik, the Czechoslovak planner, to the requirement of reducing uncertainty through the control of inputs and outputs (Owens and Sutcliffe 1972). A complex of causes, however, is evident even for an imperialism defined sensibly for a specific historical period. The so-called 'new imperialism', which is the concern here and which dates from the 1870s–80s and onwards, is attributed to economic factors by, say, Hobson (1902) and Hilferding (1910); to European diplomatic rivalries by Fieldhouse (1966) and Langer (1935); and to extreme nationalism by Hayes (1941) and Mommsen (1980).

Precisely where to draw the dividing line between imperialist episodes, however, is contentious; and more than a mere theoretical quibble in the case of the 'new imperialism'. Robinson and Gallagher (1953) argue that there is little that distinguishes the allegedly 'indifferent' mid-Victorian imperialism, when free-trade beliefs were at their height, from the 'enthusiastic' late- Victorian imperialism, when such beliefs were in decline, along with British competitiveness. According to the authors, the indifference–enthusiasm polarization leaves out too many of the facts. There were numerous additions to empire, both formal and informal, in the indifferent decades. Between 1841 and 1851 Great Britain occupied or annexed New Zealand, the Gold Coast, Labuan, Natal, the Punjab, Sind and Hong Kong. In the next 20 years British control was asserted over Berar, Oudh, Lower Burma and Kowloon, over Lagos and the neighbourhood of Sierra Leone, over Basutoland, Griqualand and the Transvaal; and new colonies were established in Queensland and British Columbia. What is more, in the supposedly laissez-faire period, before the 1870s, the economy of India was managed along the best mercantilist lines. Such continuity in nineteenth century imperialism contradicts 'those who have seen imperialism as the high stage of capitalism and the inevitable result of foreign investment . . . [in] . . . the period after the 1880s', Lenin included.

Lenin's towering influence on Marxist theorists derives from his pamphlet, *Imperialism, the Highest Stage of Capitalism,* written in 1916 in response to the outbreak of war. The academic establishment in Europe attributed the First World War mostly to the official mind. Lenin ascribed it to monopoly capitalism, the economic mainspring of imperialist rivalry:

> Railways are a summation of the basic capitalist industries: coal, iron and steel; . . . The uneven distribution of the railways, their uneven development – sums up, as it were, modern monopolist capitalism on a world-wide scale. And this summary proves that imperialist wars are absolutely inevitable under such an economic system . . . (Preface, pp. 4–5).

The economic system of monopoly capitalism is first portrayed by Lenin as being highly productive. According to a US Commission that he cites, the trusts expand their market share on the basis of scale economies and superior technology: 'Their superiority over competitors is due to the magnitude of . . . [their] . . . enterprises and their excellent technical equipment.' This leads Lenin to state: 'Competition becomes transformed into monopoly. The result is immense progress . . ..In particular, the process of technical invention and

improvement becomes socialized' (p. 24). He goes on to argue, however, that industrial capital falls prey to finance capital. He also embraces the prevailing academic view of monopoly, that it is unproductive, although he is far more cautious about this than his followers were to be:

> Certainly, the possibility of reducing cost of production and increasing profits by introducing technical improvements operates in the direction of change. But the tendency to stagnation and decay, which is characteristic of monopoly, continues to operate, and in certain branches of industry, in certain countries, for certain periods of time, it gains the upper hand (p. 119).

Stagnation, in turn, leads to the export of capital, but Lenin is vague in his explanation for why this should be so:

> The necessity for exporting capital arises from the fact that in a few countries capitalism has become 'overripe' and (owing to the backward stage of agriculture and the impoverished state of the masses) capital cannot find a field for 'profitable' investment (p. 74).

The direction of capital exports is to the backward countries:

> . . . surplus capital will be utilized . . . for the purpose of increasing profits by exporting capital abroad to the backward countries. In these backward countries profits are usually high, for capital is scarce, the price of land is relatively low, wages are low, raw materials are cheap (p. 73).

For Lenin, therefore, imperialism becomes organically inseparable from monopoly capitalism. Whereas in common usage imperialism means forced economic gain on a global scale, to Lenin it means much more. The most concise definition he gives is 'imperialism is the monopoly stage of capitalism', uniquely characterized, it should be added, by capital export.

Capital exports rose dramatically after the turn of the twentieth century. Yet neither underconsumption, as expounded by Hobson, nor a superabundance of capital, as Lenin suggested, nor a declining profit rate, a conceivable consequence of rising capital investments at home, provide particularly good explanations. Instead, Magdoff (1972) argues that in addition to the immediate causes of the sudden upsurge of capital exports (more competitors, more exporters; more tariff walls, more foreign investment to jump them), '[t]he desire and need to operate on a world scale is built into the economics of capitalism' (p. 148). Competition creates pressures for the expansion of markets. The emergence of a significant degree of concentration does not mean the end of competition. 'It does mean that competition has been raised to a new level . . ..Since capital operates on a world scale,. . . the competitive struggle among the giants for markets stretches over large sections of the globe' (p. 157). Although the scramble for colonies preceded rather than followed the rise of monopoly and capital exports, annexation was not what Lenin meant by imperialism. On the contrary, Sutcliffe states, in response to Robinson and Gallagher, 'it was a prelude to imperialism . . ..The system changed its character at the end of the century because from then on both expansion and rivalry between the major capitalist powers would have to take new forms since the chances of territorial expansion had been exhausted' (Sutcliffe 1972, p. 314).

Lenin based his analysis of imperialism on the stranglehold of finance capital, by which he meant the leading role that banks came to play in economic decision making. The financiers were perceived to have the biggest stake in imperialism and their hunger for quick returns led to economic chaos. Yet in fact after World War I finance capital decidedly took a back seat as the multinational firm grew in the US, Europe and, belatedly, England. As evidence for this, there was a shift over time away from indirect foreign investment, that is, portfolio or debt capital, to direct foreign investment, or equity capital. Roughly two-thirds of foreign investment took the form of debt capital before World War I. Thereafter, direct foreign investment became predominant, although a new type of portfolio investment rose again sharply in the late 1970s-early 1980s.

Chandler (1980) writes about the *form* that the growth of large-scale firms assumed:

> . . . modern industrial enterprise . . . grew by adding new units of production and distribution, by adding sales and purchasing offices, by adding facilities for producing raw and semi-finished materials, by obtaining . . . transportation units, and even by building research laboratories . . . (p. 397).

These new specializations of large business enterprises are the crux of Hymer's (1976) explanation for why capital exports were increasingly direct rather than indirect. According to him, the specializations that Chandler mentions – management expertise, capability in manufacturing, technology, distribution – constituted firm-specific monopolistic assets. To take full monetary advantage of them, firms exerted direct control over overseas operations, through equity ownership.

Yet foreign investment, whether direct or indirect, did not flow preponderantly to backward regions. In the interwar period and even before 1914, the main destination for overseas funds was Europe and North America. British colonies, including India, accounted for only about 20% and South America, for another 20% (Barratt Brown 1972). After 1929, the share of the advanced countries in the inflow of direct foreign investment rose even further, reaching around 75% of the total in the mid-1970s. The share was higher still for direct foreign investment in the manufacturing sector (USDC various years). Thus, while the locus of socialist revolutions was backward regions, not advanced ones, capital exports flowed increasingly to advanced regions, not backward ones. The direction of foreign investment is significant because it suggests an altogether different centre of gravity in economic activity under monopoly capitalism from the one Lenin's followers entertained.

Beginning at the turn of the century, the principal orientation of the economic activity of advanced countries was, in general, toward each other, not the backward regions. Like foreign investment, foreign trade in manufactures largely engaged the advanced countries. Their competitive struggle involved mainly invasions of each other's markets. The major contest in economic strength after World War II, between the US and Japan, barely stretched to third world shores.

Explanations other than international differences in gross profit rates must be sought for the geographical distribution of foreign investment. No definitive data exist to compare profit rates across countries. Yet profit rates are likely to have been relatively higher in backward countries, as Lenin suggests, because rates of surplus value, in the Marxist accounting sense, were higher there, at least in the 1970s in the manufacturing sector (Amsden 1981). One reason why foreign investment and trade primarily occupied the richer countries is that their per capita incomes were growing faster than the poorer regions; the newly industrializing countries excluded. The higher *level* of income in advanced countries also made them better markets. In turn, high income markets complemented the type of competition that became characteristic of monopoly capitalism. The monopolistic assets of large business enterprises were the competitive weapons. The coming of age of industrial capital witnessed an intensification of competition on the technology front. New products, new processes, new production systems constituted the razor's edge of the competitive battle, moderating the demand for protection and price-fixing cartels.

Such technology was not designed with third world domination in mind. The location of industry in the course of a product cycle from the 1950s at least through the 1980s progressed from the innovating country, to other advanced countries and only belatedly to backward regions (Vernon 1966); and then only if new discoveries did not short circuit the cycle such that production returned to the innovator's country of origin.

The monopolistic assets of large business enterprises did not all work productively, and Marxists pointed to the wasteful effects of advertising and to the ruinous effects of financial manipulation in the form of takeover waves at home and periodic, aggressive bouts of lending to the backward regions. But technological competition was the stuff out of which monopoly capitalism was made after World War II. So to equate monopoly capitalism and imperialism robs both terms of much of their meaning. The two cannot be reduced to one another.

Even if, following Stokes (1969), one attributes to Lenin what has come to be a non-'-Leninist' view, that the contestation of imperialist rivalries occurs not in the third world but in the monopolized countries themselves, then the conflation of monopoly capitalism and imperialism is still obfuscating. Whereas such rivalries engaged Europe in war at the time Lenin was

writing, they were mediated peacefully there for at least 40 years after World War II.

Nor is Lenin especially illuminating on why capital exports are the *specifica differentia* of monopoly capitalism. Was foreign investment more likely to precipitate the use of force than foreign trade? No, because trade in raw materials in the nineteenth century presupposed foreign investment. And what is the significance of the shift from indirect to direct foreign investment? Marxists have not systematically explored the answer. History teaches us that finance capital increasingly falls under the control of a few large banks, but it comprises much less differentiable products than industrial capital and, therefore, is more at the mercy of the laws of supply and demand. To prevent interest rates from falling, the banks look overseas for profitable investment outlets, and when they compete on the basis of price, they look in particular to the backward regions. The upsurge of portfolio investment in the late 1970s–early 1980s was accounted for overwhelmingly by the third world. Presumably the backward regions will become a more important locale for industrial capital as technological competition among advanced countries grows more even and product differentiation converges. Then manufacturers may be expected to locate their production facilities in lower wage, higher profit countries in order to compete better on the basis of price. That they did not do so to any significant extent before the 1980s suggests not a shortage of profitable investment outlets in advanced countries, supposedly a hallmark of monopoly, but a surplus of such outlets. Even though profit rates in the manufacturing sector were lower in the advanced countries, assuming the numbers are correct, marginal profit rates are likely to have been equal or higher, due to an outpouring of innovations.

The backward regions, however, were hardly inconsequential, to either industrial or finance capital. Certain third world raw materials, not least of all petroleum, remained critical business cost factors. The third world's debt crises undermined global monetary stability. The 'defection' of third world countries to socialism precipitated armed intervention. And, while the capital that flowed from the advanced countries to the third world throughout most of the tenure of the 'new imperialism' amounted to a mere trickle, there was a massive net transfer of surplus from the third world to the advanced countries (Bagchi 1982). Capitalism, after all, had become a world system. The relationship between imperialism and the economic development of the backward regions was the subject of as much literature as the relationship between imperialism and monopoly capitalism. Indeed, more was written on the former, because the neoclassical economists contributed; discreetly, the term imperialism never being mentioned.

Imperialism before and after World War II was quite distinct, as formal colonialism ended and large portions of Asia and Africa gained independence. One would expect economic growth in the backward regions to be quite distinct in each period as well, as a consequence of such political change. Yet, curiously, both Marxist and neoclassical economists saw continuity. In the neoclassical view, the backward regions had as good a chance to develop under colonialism as under independent rule so long as they organized their economies in the pursuit of comparative advantage. For the Marxists, underdevelopment was the expected outcome whatever the political regime, so long as the economic mechanisms of imperialism were fundamentally unaltered.

But how did these mechanisms operate? And did they remain unaltered amidst shifts in political circumstance? Schumpeter's argument, that imperialism under capitalism was a throwback to precapitalist impulses, was based on the premise that peaceful exchange was preferable to the use of force for all self-interested parties, and that ultimately reason would prevail over atavism. Yet, at minimum, force might be rational for one party to hasten another's *entry* into capitalist exchange relationships or to prevent another's *exit* into an altogether different economic system. The latter appears to have driven a good deal of US imperialism after World War II, notwithstanding the fact that the US had no precapitalist history. In the war's aftermath, American aid to Greece and Turkey limited leftist activity and the US government helped opponents of socialist and communist

candidates for office in France and Italy. Vietnam apart, the US intervened either directly through the military or covertly through the Central Intelligence Agency to halt what was perceived as socialist aggression in Greece, Iran, Guatemala, Indonesia, Lebanon, Laos, Cuba, the Congo, British Guiana, the Dominican Republic, Chile and possibly Brazil.

The onset of capitalist relations in the third world was also replete with the use of force. In many colonies where foreign enclaves were established in the nineteenth century for the purpose of producing primary products for export, population was scarce, so in retrospect 'overpopulation' cannot be held responsible for the underdevelopment that ensued. Indeed, one would have expected not underdevelopment but the onset of a 'high wage economy', given a scarcity of labour and a growing demand for labour's services in the mines and on the plantations. But wages did not rise (Myint 1964). For the neoclassical paradigm of peaceful market exchange, this constitutes a paradox. For more institutionally oriented economists, this seeming paradox was resolved with the artifice of the 'backward bending labor supply curve'. It was imagined that self-sufficient peasants who migrated to the mines and plantations offered their services with the limited purpose of obtaining only a 'target' income. If higher wages were paid, their objective would be met all the faster, with the consequence of a smaller, not larger labour supply. In fact, foreign firms in the mining and plantation sectors were faced with a decision – of whether to pay in excess of labour productivity in the short run or to coerce an adequate labour supply at a low wage rate equal to (or below) the prevailing level of productivity – and they opted for force. Colonial authorities passed legislation that indirectly compelled natives to work: but taxes were imposed that had to be paid in cash, not kind and alternative income-earning opportunities were limited through encroachments on land and restrictions on the cultivation of cash crops. The result was the onset of a 'low wage economy', that effectively channelled the 'secondary multiplier effects', of enclave production to the advanced countries and doomed the backward regions to a 'vicious circle' of poverty (Myrdal 1957; Nurkse 1953; Singer 1950).

Outright appropriation of land and labour was more blatant in the earlier than the later phases of imperialism and in some backward regions (Indonesia, the Congo) than in others (India, Latin America). But it was often possible to extract more surplus through indirect taxation and through purchase of commodities and sale of manufactures from and to the peasants. 'The British', writes Bagchi, 'may indeed be regarded as the real founders of modern neocolonialism, for both in Latin America and in India in the late nineteenth century they depended more on economic power and political influence than on direct use of political power at every stage for obtaining the lion's share of the surplus of the dominated economies' (1982, p. 78). Land taxes, payable in cash, either reduced the peasants to landless proletarians or required them to produce export crops, with little surplus to diversify in the event of unfavourable terms of trade. Free trade itself destroyed domestic manufactures, made it unprofitable to invest in anything other than export crops and impeded the growth of capitalist classes that could have challenged foreign domination. Even in the bottom of the barrel, backward regions characterized by peasant export economies with little to offer foreigners in the way of raw materials or markets (say, West Africa, Burma, Thailand and Vietnam), the functioning of the market mechanism was not devoid of coercive elements. Peasants who entered the money economy became vulnerable to international commodity price fluctuations. Foreigners, acting as monopolistic middlemen, gained the upper hand and reinvested the surplus elsewhere. Local moneylenders, who controlled credit, foreclosed on indebted peasants where land had become alienable. Railways and other infrastructure supported external rather than internal exchange, thereby discouraging domestic manufactures.

In reality, therefore, no ideal, pure market exchange between rich and poor countries existed that could be delinked neatly from imperialism. Mechanisms of coercion and mechanisms of exchange operated hand-in-hand. From the Marxist perspective it followed that imperialism was

neither atavistic nor limited merely to entry and exit to and from capitalist exchange. Rather, force was pervasive and imperialism was business as usual.

If, to varying degrees, force was pervasive in market relationships, then as force changed its colours in tandem with political change, one would expect some change in market relationships as well. Imperialism, after all, is a political phenomenon. Yet in the post-World War II period, no attempt was made by Marxists to distinguish the intrinsic from the historical effects of different economic practices on growth prospects. Instead, all intercourse with advanced countries was condemned as leading to underdevelopment, in sharp contradistinction to Marx, Engels, and even Lenin. The economic practices singled out for special opprobrium were those in which intercourse between the advanced countries and backward regions was most direct – foreign trade, foreign investment and even foreign aid. As Brenner (1977) put it, Adam Smith was turned on his head.

Yet the effect of any given economic practice on economic development clearly depended on the political setting. Aid helped Europe after World War II but seemingly hurt Bangladesh. Whereas export-led growth based on a primary product or 'staple' led to underdevelopment in the backward regions, it led to prosperity in the regions of recent settlement (Canada, Australia, New Zealand, white South Africa, white Rhodesia, etc.). Evidently there was nothing inherent in exporting that led irrevocably to either development or underdevelopment. Rather, what happened depended on local conditions. Unlike the backward regions, the regions of recent settlement retained the surplus by dint of their 'high wage economies' and reared a manufacturing sector by erecting protective tariffs. In the case of direct foreign investment, the expected gains to the 'host' country were *a priori* indeterminate. On the one hand, direct foreign investment promised a transfer of modern management techniques to backward regions. On the other hand, motivated by a wish to make use of monopolistic assets, there was nothing to insure that the multinationals would share their know-how with local managers. In fact, the outcome depended on the political conditions imposed on foreign capital; so Canada benefited far more than say, Chile, from overseas investment.

If Marxists saw foreign trade and foreign investment as dooming the third world to underdevelopment, neoclassical economists followed the same logic but arrived at an opposite conclusion: that foreign trade and foreign investment were the key to third world prosperity (Little 1982). Now this flew in the face of reality. The economies of the backward regions had long been oriented to foreign trade and foreign investment but were hardly prosperous. Two different tacks were taken to reconcile any seeming inconsistency between theory and practice. One, it was argued that the backward regions had not been sufficiently singleminded in their pursuit of free trade. They had broken faith after World War II in particular, by embracing the 'dogma of dirigisme' (whereupon, it may be added, they grew the fastest ever; Lal 1984). Two, it was argued that, in fact, the backward regions had long been growing at a fairly rapid clip, although to be sure, there were exceptions to the rule. According to Reynolds (1985): '. . . against the view that "life began in 1950," . . . the third world has a rich record of prior growth, beginning for most countries in the 1850–1914 era' (p. 4). In anticipation of the obvious objection, that developing countries are still desperately poor, Reynolds writes:

> Certainly people in Western Europe and the United States are much better off than people in Sri Lanka [the example he uses], though not as much better off as the World Bank tables suggest . . . conversion from local currencies to U.S. dollars at official exchange rates exaggerates the actual difference in consumption levels (p. 40).

Both Marxist and neoclassical analysis suffered from a failure to look beyond either the historical specificities of 'export-led exploitation' (the term is Bagchi's) or the formalism of export-led growth, as the case may be, to the underlying power structures in the backward regions. Beginning with Baran (1957), Marxists portrayed political and social life in the third world simplistically.

The state and whatever local capitalists existed were seen as corrupt puppets of advanced country powers. No scope was given to the possibility of local initiatives to mediate foreign trade, foreign investment and foreign aid to advantage. It is fair to say the neoclassical economists largely ignored local conditions in developing countries, even economic ones. When Jacob Viner (1953) delivered a lecture series in Brazil in 1950, he expressed confidence in a growth strategy based on agricultural exports. As evidence, he pointed to the correlation between high per capita incomes and agricultural exports in the regions of recent settlement, overlooking any other factors in these regions that may also have contributed to growth. The result was an inability to grasp what came to constitute a serious challenge to both theories: the economic development along capitalist lines after World War II of a handful of nations (or nation states) in East Asia, South Korea and Taiwan in particular.

The development of these countries posed a challenge to neoclassical theory because, while all the countries in question were highly oriented to trade, they were by no means committed to laissez-faire (Amsden 1985). They exerted strong centralized control over their economies. They flouted static comparative advantage and were protectionist. Their large private or public conglomerates were a mirror image of concentrations of economic power under monopoly capitalism in advanced countries. They fought force with force, as it were, in dealing with foreign capital. To say that these countries could have grown even faster had they adopted laissez-faire policies is beside the point. The development of these countries posed a challenge to Marxist theory because it wasn't supposed to happen. Such development, therefore, was preemptively dismissed. It was attributed either to a fluke – geopolitics and a superabundance of foreign aid [sic] – or repression of workers, although Engels (1878) cautions against the view that it is possible to industrialize by the gun.

The one dissenting voice among Marxists against the notion that capitalism underdevelops the third world missed the point. For Warren (1980), the problem of underdevelopment was not too much foreign capital but too little. Yet, however great the flow of foreign capital to South Korea and Taiwan (mostly, it may be noted, in the form of finance rather than industrial capital), much more accounted for development in these countries than capital per se.

The intellectual antecedents of Warren's view are traceable directly to Marx, so to suggest that Warren missed the point about economic development is also to suggest that Marx himself missed the point. Marx's point is that colonies like India were destined to develop because the capitalist system was compelled to replicate itself around the globe. With the destruction of the Asiatic mode of production, with the imposition of market relationships and with the arrival of the railroad, India would become another England (Marx and Engels 1960). Yet markets and technology alone do not make for economic development. What appears to be critical are the power relationships and institutions that unfold on their own terms to guide the accumulation process. But Marx is silent about these.

The dirigiste state stands at the opposite extreme of Marx's liberal view of the market as the engine of growth. But neither is a dirigiste regime a sufficient condition for economic development. Dirigisme and underdevelopment are both rampant in the third world. Instead, what Japan and a few South Koreans suggest is that economic development in the twentieth century hinges on a delicate relationship between the operations of the market and coercive mechanisms.

Marxists have focused on this relationship in the general case, which is the starting point for any theory of imperialism, and presuppose that markets and force are impenetrable. Yet their equation of imperialism and monopoly capitalism led them to misjudge the relationship after World War II, because imperialism was not the key to the rapid growth of the advanced countries. And their second *idée fixe,* that capitalism underdevelops the third world, led again to the relationship's misjudgement, because proof of economic development in even a handful of third world countries

deprived their theory of analytical clarity. Nonetheless, to operate with the world view of the neoclassicists – of a separation between markets and power – is to deny the very existence of imperialism and to forego the conceptual tools to analyse it.

See Also

- ▶ Colonialism
- ▶ Colonies
- ▶ Hobson, John Atkinson (1858–1940)
- ▶ Lenin, Vladimir Ilyich [Ulyanov] (1870–1924)
- ▶ Nationalism
- ▶ Periphery
- ▶ Unequal Exchange

Bibliography

Amin, S. 1976. *Unequal development*. New York: Monthly Review Press.

Amsden, A.H. September, 1981. An international comparison of the rate of surplus value in manufacturing industries. *Cambridge Journal of Economics* 5(3): 229–49.

Amsden, A.H. 1985. The state and Taiwan's economic development. In *Bringing the state back in*, ed. P. Evans et al. Cambridge: Cambridge University Press.

Bagchi, A. 1982. *The political economy of under development*. Cambridge: Cambridge University Press.

Baran, P.A. 1957. *The political economy of growth*. New York: Monthly Review Press.

Barratt Brown, M. 1963. *After imperialism.* London: Merlin Press, 1970.

Barratt Brown, M. 1972. A critique of Marxist theories of imperialism. In Owen and Sutcliffe (1972).

Brenner, R. 1977. The origins of capitalist development: A critique of neo-Smithian Marxism. *New Left Review* 104, July–August, 25–92.

Bukharin, N. 1914. *Imperialism and world economy.* New York: Monthly Review Press. 1972.

Chandler, A. August, 1980. The growth of the transnational industrial firm in the United States and the United Kingdom: A comparative analysis. *Economic History Review* 33: 396–410.

Engels, F. 1878. *Anti-Dühring*. New York: International Publishers.

Fieldhouse, K.D. 1966. *The colonial empires: A comparative study from the eighteenth century.* London: Weidenfeld & Nicolson.

Hayes, C.J.H. 1941. *A generation of materialism 1871–1900*. New York: Harper.

Hilferding, R. 1910. *Finance capital: A study of the latest phase of capitalist development.* London: Routledge & Kegan Paul, 1981.

Hobson, J.A. 1902. *Imperialism: A study.* London: Allen & Unwin, 1938.

Hymer, S. 1976. *The international operations of national firms: A study of direct foreign investment*. Cambridge, MA: MIT Press.

Lal, D. 1984. *The poverty of 'Development of economics'*. London: Institute of Economic Affairs.

Langer, W.L. 1935. *The diplomacy of imperialism 1890–1902*, 2nd ed. New York: Knopf.

Lenin, V.I. 1973. *Imperialism, the highest stage of capitalism*. Peking: Foreign Language Press.

Little, I. 1982. *Economic development: Theory, policy, and international relations*. New York: Basic Books.

Luxemburg, R. 1913. *The accumulation of capital.* Trans. Agnes Schwarzchild. London: Routledge & Kegan Paul, 1951.

Magdoff, H. 1972. Imperialism without colonies. In Owen and Sutcliffe (1972).

Marx, K., and F. Engels. 1960. *On colonialism*. London: Lawrence & Wishart.

Mommsen, W. 1980. *Theories of imperialism.* Trans. P.S. Falla. New York: Random House.

Myint, H. 1964. *The economics of the developing countries*. New York: Praeger.

Myrdal, G. 1957. *Rich lands and poor: The road to world prosperity.* New York: Harper.

Nurkse, R. 1953. *Problems of capital formation in underdeveloped countries*. New York: Oxford University Press.

Owen, R., and B. Sutcliffe (eds.). 1972. *Studies in the theory of imperialism*. New York: Longmans.

Reynolds, L.G. 1985. *Economic growth in the third world, 1850–1980*. New Haven: Yale University Press.

Robinson, R., and Gallagher, R. 1953. The imperialism of free trade. *Economic History Review,* 2nd Series 6(1): 1–15.

Schumpeter, J. 1919. *Imperialism and social classes.* Trans. New York: Augustus M. Kelley.

Singer, H.W. 1950. The distribution of gains between investing and borrowing countries. *American Economic Review, Papers and Proceedings* 40, May, 473–85.

Sutcliffe, B. 1972. Conclusion. In Owen and Sutcliffe (1972).

Stokes, E. 1969. Late nineteenth-century colonial expansion and the attack on the theory of economic imperialism: A case of mistaken identity? *Historical Journal* 12: 285–301.

United States Department of Commerce (USDC). (Various Years.) *Survey of Current Business.*

Vernon, R. 1966. International investment and international trade in the product cycle. *Quarterly Journal of Economics* 80, May, 190–207.

Viner, J. 1953. *International trade and economic development*. Oxford: Clarendon.

Warren, B. 1980. *Imperialism: Pioneer of capitalism*. London: Verso.

Implicit Contracts

Costas Azariadis

JEL Classifications
J41

An implicit contract is a theoretical construct meant to describe complex agreements, written and tacit, between employers and employees, which govern the exchange of labour services when various types of job-specific investments inhibit labour mobility and opportunities to shed risk are limited by imperfectly developed markets for contingent claims. This construct differs from the more familiar one of a neoclassical labour exchange in emphasizing a trading process, frequently over a long period of time, between two *specific* economic units (say a worker and a firm, union and management, and so on) rather than the impersonal, and often instantaneous, market process in which wages decentralize and coordinate the actions of labour suppliers and labour demanders.

Adam Smith's exposition of occupational wage differentials (1776, book I, ch. 10) recognized very early the idiosyncratic nature of the labour market and, in particular, that employment risk affected wages in various occupations. Since then economists have accumulated many facts, raw or stylized, which are best understood if one abandons the traditional view that the shadow price of labour is simply the wage rate. Prominent among explananda are the widespread use of temporary layoffs as a means of regulating the volume of employment (Feldstein 1975); the continuity of jobs by many primary wage earners (Hall 1982); the collective bargaining tradition of leaving the volume of employment at the discretion of management while predetermining money wage rates two or three years in advance.

To these, one must add certain 'impressions' or softer facts about the labour market which arise from the central role labour services possess in macroeconomic models. There is, indeed, among macroeconomists a shared impression (Hall 1980) that, over a typical business cycle, average real compensation per hour fluctuates considerably less than does the marginal revenue-product of labour or, for that matter, the total volume of employment.

One consequence is that wage and price rigidity are among the key assumptions of Keynesian macroeconomics, both in the Hicksian IS-LM framework and in the concept of quantity-constrained equilibrium originally developed by Clower (1965) and formalized by Bénassy (1975) and Drèze (1975). Another is the overwhelming importance of words like 'jobs' and 'unemployment', both in our colloquial vocabulary and in the specialized lexicon of economics. In particular, 'involuntary unemployment' is for many academic economists the *sine qua non* of modern macroeconomics.

The technically minded reader will find many of these issues surveyed in a number of specialized papers of which the most recent are Hart (1983) and Rosen (1985).

Wages and Employment

The earliest literature on implicit contracts exploits an insight of Frank Knight (1921), who argued that inherently 'confident and venturesome' entrepreneurs will offer to relieve their employees of some market risks in return for the right to make allocative decisions. The formal development of this idea began with three independently written papers by Baily (1974), Gordon (1974), and Azariadis (1975), motivated by the seeming puzzle of layoffs. In an unusual coincidence, all three authors took the employment relation not simply as a sequential spot exchange of labour services for money, but as a more complicated long-term attachment; labour services are traded in part for an insurance contract that protects workers from random, publicly observed fluctuations in their marginal revenue-product. The idea was that workers could purchase insurance only from their employers, not from third parties.

Risk-averse workers deal with risk-neutral entrepreneurs who head firms consisting of three departments: a production department that purchases labour services and credits each worker with his marginal revenue-product (MRPL); an insurance department that sells actuarially fair policies and, depending on the state of nature, credits the worker with a net insurance indemnity (NII) or debits him with a net insurance premium; and an accounting department that pays each employed worker a wage, w, with the property that w = MRPL + NII in every state of nature.

Favourable states of nature are associated with high values of MRPL; in these the net indemnity is negative and wage falls short of the MRPL. Adverse states of nature correspond to low values of MRPL, to positive net insurance indemnities, and to wages in excess of MRPL. An implicit contract is then a complete description, made before the state of nature becomes known, of the labour services to be rendered unto the firm in each state of nature, and of the corresponding payments to be delivered to the worker. The contract is implementable if we assume the state of nature is as easily verifiable as events are in a normal insurance contract.

An immediate consequence of this framework is that wages are disengaged from the marginal revenue-product of labour. In fact, if the amount of labour performed by employed workers per unit time is fixed institutionally, then each worker's consumption is proportional to the wage rate; an actuarially fair insurance policy should make this consumption independent of the MRPL by stabilizing the purchasing power of wages over states of nature. Therefore, the real wage rate is rigid.

In traditional macroeconomic models of course, wage rigidity by itself is sufficient to cause unemployment: if wages do not adjust for some reason, then neither does the demand for labour. The argument does not carry over to implicit contracts because of the very separation between wages and the marginal revenue product of labour. A complete theory of unemployment must explain why layoffs are preferred to work-sharing in adverse states of nature, and why laid-off workers are worse off than their employed colleagues.

This is not a simple task if one thinks of implicit contracts as ordinary, explicit, timeless insurance contracts between risk-averse workers and risk-neutral entrepreneurs. All contracts of this type would share a basic property of optimum insurance schemes; namely, keeping the worker's marginal utility of consumption independent of all random, publicly observed events - including such events as 'employment' or 'unemployment'.

To explain layoff unemployment, we need to distort or complicate the insurance contract in some significant way. A distortion that was noted early in the implicit contract literature is the dole. In an extremely adverse state of nature, the flow of insurance indemnities to workers can become a substantial drain on profit; one way to staunch losses is to place the burden of insurance on an outside party, the dole.

The practice of layoffs is simply the administrative counterpart of this insurance-shifting manoeuvre; workers consent in advance that some of them may be separated from their jobs in order to become eligible for unemployment insurance (UI) payments from an outside public agency. Furthermore, no worker will contract his labour unless the expected value (utility) of the total package, taken over all possible states of nature, exceeds the value of being on the dole in every state. This means, in turn that employed workers receive a wage in excess of UI payments and are therefore to be envied by their laid-off colleagues - a situation that many economists would call 'involuntary unemployment'.

The fact that laid-off workers would gladly exchange places with their employed colleagues is not in itself sufficient to establish a misallocation of resources. After all, accident victims may very well envy more fortunate individuals without any implication that the insurance industry works poorly. Layoffs, by themselves, could be no more than the luck of the draw unless we can demonstrate that they constitute, in some sense, socially inefficient underemployment. This is clearly impossible within the Walras-Arrow-Debreu model; and it is for this reason that the early literature on contracts turned to institutions like the dole in order to explain layoff unemployment.

Private Information

One fundamental departure from the Walrasian paradigm that received much attention in the early 1980s was a weakening of the information assumptions: information becomes 'private' or 'asymmetric', which simply means that not everyone is equally informed about the relevant state of nature. This is a perfectly sensible observation, for what justifies the trading of implicit contracts in the first place is that third parties simply are not as well informed about someone's income or employment status as is his employer; the employer, in turn, may be less informed about an employee's non-labour income and job opportunities than is the worker himself.

The thread was picked up by a number of authors who studied the properties of wages and employment for two main cases: in the first, entrepreneurs possess superior information about labour demand (Hall and Lilien 1979; Grossman and Hart 1981; Azariadis 1983; Farmer 1984); in the second case, workers possess superior information about labour supply, as in Cooper (1983). Suppose, for instance, that wages and employment do not depend on the unobservable true state of nature but on what the better informed contractant (say, the employer) *announces* that state to be. The question now becomes how to design contracts that reward entrepreneurs who tell the truth and punish those who lie.

One desirable property of contracts is that the truth should be the value- maximizing strategy for firms: truth-telling ought to be consistent with equality between the marginal cost and the marginal revenue-product of labour. Furthermore, entrepreneurs who misrepresent actual conditions should be punished, say, for knowingly under-reporting demand.

Under-reporting demand does turn out to be a problem in contracts that permit employers to slash both workforce and the wage bill when demand is slack, and do it in such a manner as to reduce cost more than revenue. To avoid this temptation, a properly designed contract specifies a highly variable pattern of employment over states of nature; that is, one in which employment is below what is socially optimal and the marginal product of labour is correspondingly above the marginal rate of substitution between consumption and leisure. It is in this sense that asymmetric information is said to result in socially inefficient underemployment or unemployment.

What relation is there between the layoffs we all know and the inefficient underemployment of a model economy that suffers from asymmetric information? To go from the latter to the former, one must understand first why layoffs are a more common means of reducing employment than is work-sharing. Second, a general equilibrium picture of underemployment would require an explanation of why underemployed (or unemployed) individuals are not hired by other employers. Third, and most important, the unemployment found in this private-information story is a response to private, firm-specific risk; most economists, however, consider the unemployment observed in market economies to be a reaction to social risks, especially to business cycles set in motion by aggregate demand disturbances. Unless one intends to make the far-fetched claim that the general public is unaware of, or cannot observe, whatever disturbances set off business cycles (such as changes in government consumption, money supply or consumer confidence), does it not appear that information-based unemployment simply describes the behaviour of an isolated firm?

The answer is not obvious. Note, however, that in order to have an inefficient volume of equilibrium employment, it is sufficient that *some but not all* information be private. In fact, it is not difficult to imagine general equilibrium extensions of the work we are discussing that would include both public and private information. Such extensions will be useful, especially if they manage to establish a firm link between inefficient underemployment and extreme values of some publicly observed aggregate disturbance.

Empirical Implications

Whether information is publicly shared or in the private domain, wages in implicit contracts do not merely reflect the marginal product of labour or

the workers' marginal rate of substitution between consumption and leisure, as they might in more conventional theories. The empirical implications of this insight are just being worked out, and they seem to be quite considerable. At the most aggregative level, one can make sense of the oft-verified fact (Neftci 1978) that hourly wages in manufacturing show little cyclical variability and are best described as a random walk.

In fact, it seems preferable to have empirical investigations of this sort at a less aggregated level. Aggregate studies are victims of selection bias: they fail to capture changes in the composition of output or of the labour force, which are themselves sufficient to induce substantial cyclical movement in economy-wide wages even if the business cycle does not affect the real wage of any skill grade in any industry.

Consider, for instance, a fictitious economy with homogeneous labour in which almost all industries experience little cyclical fluctuation except one, the quadindustry, which is thoroughly buffeted by the business cycle. If labour mobility is good across industries, quad workers will suffer more layoffs and enjoy a wage higher than elsewhere whenever they are employed. The economy-wide average wage will vary procyclically.

Another phenomenon accounted for naturally by implicit contracts is the behaviour of occupational wage differentials (that is of the unskilled-to-skilled wage ratio). These have shown a definite countercyclical tendency, widening in contractions and narrowing in booms, both in the United States and in the UK.

To see why, suppose that we drop the postulate of labour homogeneity in the economy just described and admit two skill grades. For simplicity, assume that the cycle is of such amplitude that there is no unemployment outside the quad industry, while unemployment in the quad industry falls solely on common labourers. These workers are thus the only group in the economy to suffer layoffs; in return they receive a wage above that of common workers outside the quad industry and below that of skilled workers - in the quad industry or out. As the cycle unfolds, then, the economy-wide wage average for craftsmen remains unaltered, the one for labourers changes procyclically, and occupational wage differentials follow a countercyclical pattern.

Intertemporal labour supply models of the type pioneered by Lucas and Rapping (1969) are another area that may in the future make fruitful use of implicit contracts. Econometric work on intertemporal labour substitution identifies the preferences of a 'typical' working household from time-series data on wages and salaries. The outcome is invariably an estimate of the wage-elasticity of labour supply that is so low as to be inconsistent with time-series data on employment (Kydland and Prescott 1982). In other words, someone who believes that the wage rate represents an important conditioning factor for labour supply and demand will find that wage rates do not vary sufficiently over the business cycle to account for observed fluctuations in employment.

Employment in an implicit contract, however, reflects the underlying value of labour's marginal revenue-product, whereas wages are smoothed averages of the MRPL over time or states of nature. Small fluctuations in contract wages are in principle consistent with substantial variations in contract employment; whether these are mutually consistent *in practice* remains to be seen from empirical work.

Macroeconomic Aspects

From empirical labour economics we turn to the macroeconomic issues that provided the original impetus for the development of implicit contracts. Unemployment, says this theory, is the result of differential information: a credible signal from employers to employees that product demand is slackening, or one from employees to employers that job opportunities are really better elsewhere.

Newer ideas that seem to be building on this basic piece of intuition are outlined later in this article. But whatever progress we have made towards understanding fluctuations in employment has not dispelled the dense fog that still shrouds the issue for wage rigidity. All we have to go on is the early result of Martin Baily that insurance makes the wage rate less variable than it

otherwise might be. This stickiness, however, is a property of the *real* rather than the nominal wage rate, and it is the latter that is assumed to be rigid in Keynesian macroeconomics.

Rigidity, of course, does not necessarily imply complete time-invariance, nor does it require money wages to change less frequently than other prices; it is simply an information-processing failure. The standard procedure in collective bargains, for instance, is to predetermine money wages several years in advance; more often than not those wages are invariant to any information that may accumulate over the duration of the contract. Only in exceptional circumstances are money wages in the United States allowed to reflect *any* contemporaneous developments in the cost of living (indexation) or in the profitability of the employer (bankruptcy).

The mystery of wage rigidity is then the failure of contracts to set money wages as *functions* of publicly available information that is obviously relevant to the welfare of all parties. Why does the wage-setting process choose to ignore this information? One answer is transaction costs and/or bounded rationality: contracts are cheaper to evaluate and implement when they are defined by a few simple numbers rather than by complicated rules that condition employment or wages on contingent events. Another possibility is to exploit the great multiplicity of equilibria that is typical of economies with missing securities markets (Azariadis and Cooper 1985). One of these equilibria features predetermined prices and wages, while employment and other quantity variables adjust fully to short-term disturbances. Wage rigidity here is like a Nash equilibrium: it is the best response of a firm in a labour market in which the wages paid by all other firms fail to reflect new information instantaneously.

Implementation

An implicit contract is formally defined as a collection of schedules describing how the terms of employment for one person or group of persons change in response to unexpected changes in the economic environment. What brings contractants together? How detailed are their agreements? And what mechanisms are there to enforce such agreements once they are reached? After an initial stage of fairly rapid development, research is returning to these elementary questions as if trying to clarify the axiomatic basis of the underlying theory.

What brings potential contractants together is the opportunity jointly to reap substantial returns on investments peculiar to their relationship. The idea is apparent in Becker's theory of specific human capital (1964) and in Williamson's hypothesis (1979) of physical assets that are specific to a given supplier–customer pair. To reap any returns, contractants must wed themselves to one partner, forsaking all others, for some period of time. Maintaining such a special relationship involves the transactions costs of creating an idiosyncratic asset, as well as an implicit contract; that is, a number of rules that define how the partners have decided to share the returns in various possible future circumstances.

There are, of course, circumstances that are not explicitly covered, either because they are not observable at reasonable cost or because contractants think of them as unlikely or unworthy of note. Irrespective of the possible events that are covered and of the prior rules that govern the distribution of returns to shared investments, all contractants are required to bear risk and to subordinate their short– term interest to longer-term considerations.

Workers, for instance, suffer layoffs in recessions while firms hoard labour in order to preserve a long-term relationship. What mechanisms keep contractants together in adverse circumstances?

One mechanism – studied extensively by Radner (1981), Townsend (1982) and others – is reputation: if somebody deviates from the terms of the contract, the deviation becomes widely known, and the deviant finds it difficult to locate trading partners in the future. That works well if the time horizon is fairly long or the future is fairly important relative to the present; reputations are likely to be important for firms, less so for workers.

Another method of enforcement is by a third party: a monitor, arbitrator or court of law. In order for a third party to enforce a contract, it has to be

able to observe all the prices and all the quantities specified in it – the employment status, hours worked and wage rate of every worker. That is an unreasonably large informational burden to place on someone who is outside the special relationship called a contract. Outsiders can be expected to observe at low cost only certain aggregates or averages, but not very much in the way of idiosyncratic detail.

How does one design and enforce contracts when outsiders are poorly informed about the trades among contractants? According to Hölmstrom (1983) and Bull (1986), self-interest will enforce contracts that third parties are not sufficiently informed to implement.

In particular, workers will put in the required amount of effort on the job, not because effort can be ascertained easily by an outside arbitrator but rather because they know that their wages and speed of promotion depend on performance. And employers will be careful not to break even the most implicit of their commitments if doing so will compromise their ability to attract workers in the future. As of this writing, the design of self-enforcing contracts seems to be the central theoretical problem in the field of implicit contracts.

See Also

- ▶ Labour Economics
- ▶ Layoffs

Bibliography

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83: 1183–1202.

Azariadis, C. 1983. Employment with asymmetric information. *Quarterly Journal of Economics* 98 (Supplement): 157–172.

Azariadis, C., and R. Cooper. 1985. Nominal wage-price rigidity as a rational expectations equilibrium. *American Economic Review, Papers and Proceedings* 75: 31–35.

Baily, M. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41: 37–50.

Becker, G. 1964. *Human capital*. New York: Columbia University Press.

Bénassy, J.-P. 1975. Neo-Keynesian disequilibrium in a monetary economy. *Review of Economic Studies* 42: 502–523.

Bull, C. 1986. The existence of self-enforcing implicit contracts. *Quarterly Journal of Economics.*

Clower, R. 1965. The Keynesian counter-revolution: A theoretical appraisal. In *The theory of interest rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Cooper, R. 1983. A note on overemployment/underemployment in labor contracts under asymmetric information. *Economics Letters* 12: 81–87.

Drèze, J. 1975. Existence of an equilibrium under price rigidity and quantity rationing. *International Economic Review* 16: 301–320.

Farmer, R. 1984. A new theory of aggregate supply. *American Economic Review* 74: 920–930.

Feldstein, M. 1975. The importance of temporary layoffs: An empirical analysis. *Brookings Papers on Economic Activity* 3: 725–744.

Gordon, D.F. 1974. A neoclassical theory of Keynesian unemployment. *Economic Inquiry* 12: 431–449.

Grossman, S., and O. Hart. 1981. Implicit contracts, moral hazard and unemployment. *American Economic Review, Papers and Proceedings* 71: 301–307.

Hall, R. 1980. Employment fluctuations and wage rigidity. *Brookings Papers on Economic Activity* 1: 91–124.

Hall, R. 1982. The importance of lifetime jobs in the US economy. *American Economic Review, Papers and Proceedings* 72: 716–724.

Hall, R., and D. Lilien. 1979. Efficient wage bargains under uncertain supply and demand. *American Economic Review* 69: 868–879.

Hart, O. 1983. Optimal labour contracts under asymmetric information: An introduction. *Review of Economic Studies* 50: 3–35.

Hölmstrom, B. 1983. Equilibrium long-term labor contracts. *Quarterly Journal of Economics* 98 (Supplement): 23–54.

Knight, F. 1921. *Risk, uncertainty and profit*. Boston: Houghton Mifflin.

Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Lucas, R., and L. Rapping. 1969. Real wages, employment and inflation. *Journal of Political Economy* 77: 721–754.

Neftci, S. 1978. A time-series analysis of the real wages–employment relationship. *Journal of Political Economy* 86: 281–291.

Radner, R. 1981. Monitoring cooperative agreements in a repeated principal–agent relationship. *Econometrica* 49: 1127–1148.

Rosen, S. 1985. Implicit contracts: A survey. *Journal of Economic Literature* 23: 1144–1175.

Smith, A. 1776. *The wealth of nations*, 1970. London: Pelican Books.

Townsend, R. 1982. Optimal multiperiod contracts and the gain from enduring relationships under private information. *Journal of Political Economy* 90: 1166–1186.

Williamson, O. 1979. Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics* 22: 233–261.

Import Substitution and Export-Led Growth

John Eatwell

In an economy in which expansion is limited by a balance of payments constraint, action must be taken either to boost exports or to limit imports. This truism takes on an added dimension when the trade strategy adopted is part of a general development strategy. In these circumstances the evaluation of any particular trade strategy must include not only the implications for the allocation of resources, but also the consequences for the rate of accumulation and of technological progress.

In the 1950s and early 1960s, the years of the dollar shortage, the balance of payments constrained industrial countries adopted quite different trade and industrial strategies. West Germany pursued a strategy of export expansion by means of an undervalued Deutschmark and subsidies to export industries. With world trade in manufactures growing rapidly, and West Germany's share of that trade growing too, the rapid growth of manufactured exports provided the foundation for domestic expansion (Shonfield 1963). Italy pursued a similar strategy by means of regular devaluation of the lira, devaluations often being associated with a surplus on Italy's current account.

These two examples of export-led growth contrast markedly with the strategies adopted by France and by Japan. Both countries vigorously protected their home markets, using industrial expansion within the home market as a springboard for the capture of export markets. The rationale behind this policy of import substitution was spelt out by Vice-Minister Ojimi, of the Japanese Ministry of International Trade and Industry:

> After the war, Japan's first exports consisted of such things as toys or other miscellaneous merchandise and low-quality textile products. Should Japan have entrusted its future, according to the theory of comparative advantage, to these industries characterized by intensive use of labour? That would perhaps be rational advice for a country with a small population of 5 or 10 million. But Japan has a large population. If the Japanese economy had adopted the simple doctrine of free trade and had chosen to specialise in this kind of industry, it would almost permanently have been unable to break away from the Asian pattern of stagnation and poverty . . .
>
> The Ministry of International Trade and Industry decided to establish in Japan industries which require intensive employment of capital and technology, industries that in consideration of comparative cost should be the most inappropriate for Japan, industries such as steel, oil refining, petrochemicals, automobiles, industrial machinery of all sorts, and electronics, including electronic computers. From a short-run, static viewpoint, encouragement of such industries would seem to be in conflict with economic rationalism. But from a long-range viewpoint, these are precisely in industries where income elasticity of demand is high, technological progress is rapid, and labour productivity rises fast . . . (Ojimi 1970).

Ojimi's argument encapsulates the dispute over import substitution or export-led growth as development strategies. The orthodox theory of international trade suggests that resources are most efficiently allocated in a regime of free trade. Efficient development would therefore require the adoption of free trade, with variation in exchange rates being used as the means of balancing trade.

This argument rests on a number of strong assumptions, in particular the assumptions that all countries have access to the same technologies, that factor markets clear (labour is fully employed), and that all countries have equal access to all markets – including equal access to all financial markets. If these, and other well-known assumptions, are not fulfilled, then the argument for free trade *on these grounds* no longer stands, and is superseded by the uncertainties of the second best.

Rejection of arguments for the efficiency of the price mechanism, for example on Keynesian grounds, also lead to the rejection of the efficiency of free trade. It was Keynesian arguments that underpinned the so-called ECLA strategy for structural change in Latin America. If expansion of domestic demand could be prevented, by protective measures, from leaking abroad then savings and fiscal revenues at home would finance

domestic investment and government expenditure. Moreover, the profitability of protected domestic production would encourage further investment. The process of expansion would be self-sustaining.

The application of import-substitution strategies in Latin America in the 1950s met initially with considerable success. Output of domestically produced manufactured goods grew rapidly, as did industrial employment. Later the policy fell into disrepute. It was argued that import substitution took place primarily in 'soft' consumer goods industries, whereas investment goods continued to be imported. Hence after the early growth associated with import substitution in consumer goods, growth was once again constrained by the necessity of importing machinery. Moreover, it was argued that protected domestic industry was relatively inefficient, and unable to compete on world markets. These matters are the subject of considerable dispute, particularly as they involve not only questions of economic efficiency, but also issues of national sovereignty, since the IMF has responded to the difficulties in which some Latin American countries have found themselves by demanding the removal of the trade protection on which the earlier development strategy was based.

These criticisms of import substitution extend beyond the traditional case for free trade, to consideration of the implication of different trade strategies for structural development and technological change. It was on exactly these grounds that Ojimi sought to justify Japan's strategy of import substitution. The Japanese case suggests that the traditional dichotomy between import substitution and export-led growth is invalid. Whilst Japanese industry was developed within a rapidly growing and protected home market, that growth proved to be springboard for expansion into world markets. Exports were domestic-growth led.

The performance of the successful Japanese (and French) examples of import substitution, and the problems encountered in Latin America, cannot be evaluated using static conceptions of allocative efficiency. Success (and lack of it) have clearly been associated with technological progress and industrial modernization. The case for free trade must be made on the ground that it encourages the most rapid adoption of the new techniques which determine competitive advantage.

Nicholas Kaldor's version of Verdoorn's Law (Kaldor 1966), whereby it is argued that the rate of productivity growth in manufacturing industry is a function of the rate of growth of demand for manufactured products, provides a framework within which trade strategies may be evaluated (see, for example, Brailovsky 1981).

The growth of demand for a country's manufactures is a function of the rate of growth of its home market, the rate of growth of its export markets, and the rate of change of its share of those markets. Changing market shares is a slow and uncertain business. It is growth of markets which is the major determinant of growth of demand. Since all countries are competing for shares of (roughly) the same export market, it is growth of the home market which typically differentiates the growth of demand for the manufactures of one country from those of another. This would suggest that manipulation of growth of the home market, using whatever means are necessary to relax the balance of payments constraint, is the most efficient development strategy.

However, the Verdoorn argument does not encompass the scale of productivity response to any given growth of demand. The implementation of industrial policies which both ensure that the expansion of industrial structure is 'balanced', and hence not overly dependent on imports, and directs demand toward those sectors which have both greatest competitive potential and which have the highest ratio of domestic value-added to import content, are more likely produce a greater response than if these issues are neglected.

The efficiency of any given trade strategy is not independent of the performance of the world economy as a whole. All countries cannot achieve export-led growth at once. Moreover, the success of the West Germany recovery strategy was undoubtedly enhanced by the fact that it was implemented in a period of rapid growth in world trade. In an era in which world trade is expanding relatively slowly, reliance on export

demand is unlikely to prove a successful foundation for rapid growth of demand and hence for rapid technological progress.

See Also

- Autarky
- Effective Protection
- Free Trade and Protection
- Immiserizing Growth
- Infant Industry
- Quotas and Tariffs
- Vent for Surplus

Bibliography

Brailovsky, V. 1981. Industrialisation and oil in Mexico: A long-term perspective. In *Oil or industry?* ed. T. Barker and V. Brailovsky. London: Academic.

Kaldor, N. 1966. *The causes of the slow rate of growth of the UK*. Cambridge: Cambridge University Press.

Ojimi, V. 1970. Japan's industrialisation strategy. In *Japanese industrial policy*, ed. OECD. Paris: OECD.

Shonfield, A. 1963. *Modern capitalism*. Oxford: Oxford University Press.

Impulse Response Function

Helmut Lütkepohl

Abstract

Impulse response functions are useful for studying the interactions between variables in a vector autoregressive model. They represent the reactions of the variables to shocks hitting the system. It is often not clear, however, which shocks are relevant for studying specific economic problems. Therefore structural information has to be used to specify meaningful shocks. Structural vector autoregressive models and the estimation of impulse responses are discussed and extensions to models with cointegrated variables or nonlinear features are considered.

Keywords

Bayesian methods; Bootstrap; Cointegrated variables; Cointegration; Conditional moment profiles; Dynamic multipliers; Forecast error impulse responses; Generalized impulse responses; Impulse response functions; Integrated variables; Least squares; Linear models; Maximum likelihood; Nonlinear time series models; Orthogonalized impulse responses; Simultaneous equations models; Structural impulse responses; Structural vector autoregressions; Vector autoregressions; Wold causal ordering; Wold moving average

JEL Classifications

C32

Sims (1980) questioned the way classical simultaneous equations models were specified and identified. He argued in particular that the exogeneity assumptions for some of the variables are often problematic. As an alternative he advocated the use of vector autoregressive (VAR) models for macroeconometric analysis. These models have the form

$$y_t = A_i y_{t-1} + \cdots + A_p y_{t-p} + u_t,$$

where $y_t = (y_{1t},\ldots,y_{Kt})'$ (the prime denotes the transpose) is a vector of K observed variables of interest, the A_i's are (K × K) parameter matrices, p is the lag order and u_t is an error process which is assumed to be white noise with zero mean, that is, $E(u_t) = 0$, the covariance matrix, $E\left(u_t u_t'\right) = \Sigma_u$ is time invariant and the u_t's are serially uncorrelated or independent. There are usually also deterministic terms such as constants, seasonal dummies or polynomial trends. These terms are neglected here because they are not of interest in what follows. The relations between the variables in a VAR model are difficult to see directly from the parameter matrices. Therefore, *impulse response functions* have been proposed as tools for interpreting VAR models.

AVAR model can be written more compactly as $A(L)y_t = u_t$, where the lag or back-shift

operator L is defined such that $Ly_t = y_{t-1}$ and $A(L) = I_K - A_1L - \cdots - A_pL^p$ is a matrix polynomial in the lag operator. If the polynomial in z defined by $\det A(z)$ has all its roots outside the complex unit circle, the process is stationary and has a Wold moving average (MA) representation

$$y_t = A(L)^{-1}u_t = u_t + \sum_{i=1}^{\infty} \Phi_i u_{t-i}. \qquad (1)$$

In this framework impulse response analysis may be based on the counterfactual experiment of tracing the marginal effect of a shock to one variable through the system by setting one component of u_t to one and all other components to zero and evaluating the responses of the y_t's to such an impulse as time goes by. These impulse responses are just the elements of the Φ_i matrices. Because the u_t's are the one-step ahead forecast errors of the system, the resulting functions are sometimes referred to as *forecast error impulse responses* (for example, Lütkepohl 2005, section 2.3.2).

Such a counterfactual experiment may not properly reflect the actual responses of an economic system of interest because the components of u_t are instantaneously correlated, that is, Σ_u may not be a diagonal matrix. In that case, forecast error impulses are just not the kinds of impulses that occur in practice, because an impulse in one variable is likely to be accompanied by an impulse in another variable and should not be considered in isolation. Therefore, *orthogonalized impulse responses* are often considered in this context. They are obtained from (1) by choosing some matrix B such that $BB' = \Sigma_u$ or such that $B^{-1}\Sigma_u B'^{-1}$ is a diagonal matrix and defining $\varepsilon_t = B^{-1}u_t$. Substituting in (1) gives

$$y_t = B\varepsilon_t + \sum_{i=1}^{\infty} \Theta_i \varepsilon_{t-i}, \qquad (2)$$

where $\Theta i = \Phi_i B$, $i = 1, 2, \ldots$ The ε_t's have a diagonal or even a unit covariance matrix and are hence contemporaneously uncorrelated (orthogonal). Thus, ε_t shocks may give a more realistic picture of the reactions of the system. The problem is, however, that the matrix B is not unique and many different orthogonal shocks exist. Thus, identifying restrictions based on non-sample information are necessary to find the unique impulses of interest which represent the actual responses of the system to shocks that occur in practice. These considerations have led to what is known as *structural VAR (SVAR) models* and *structural impulse responses*.

SVAR Models

Various types of restrictions have been considered for identifying the structural innovations or, equivalently, for finding a unique or at least locally unique B matrix. For example, using a triangular B matrix obtained from a Choleski decomposition of Σ_u is quite popular (for example, Sims 1980; Christiano et al. 1996). Choosing a lower-triangular matrix amounts to setting up a recursive system with a so-called *Wold causal ordering* of the variables. One possible interpretation is that an impulse in the first variable can have an instantaneous impact on all other variables as well, whereas an impulse in the second variable can also have an instantaneous effect on the third to last variables but not on the first one, and so on. Because such a causal ordering is sometimes difficult to defend, other types of restrictions have also been proposed. Examples are:

1. Instantaneous effects of some shocks on certain variables may be ruled out. In other words, zero restrictions are placed on B just as in the Choleski decomposition approach. The zero restrictions do not have to result in a triangular B matrix, however.
2. Identification is achieved by imposing restrictions on the instantaneous relations of the variables. In this case a structural form model of the type $A_0y_t = A_1y_{t-1} + \cdots + A_py_{t-p} + \varepsilon_t$ may be considered and typically linear restrictions are imposed on A_0. Usually the elements on the main diagonal of A_0 will be normalized to unity. The restrictions on A_0 imply restrictions

for $B = A_0^{-1}$. For example, if A_0 is triangular, then so is B.

3. It is also possible to set up a model in the form $A_0 y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + B\varepsilon_t$ and impose restrictions on both A_0 and B to identify structural shocks. Combining restrictions on B with those on the instantaneous effects on the observed variables results in the so-called AB-model of Amisano and Giannini (1997).
4. There may be prior information on the long-run effects of some shocks. In this case restrictions may be placed on $B + \sum_{i=1}^{\infty} \Theta_i = A(1)^{-1}B$ (for example, Blanchard and Quah 1989). For instance, demand shocks may be assumed to have no accumulated long-run effects on some variable (in their case output). In fact, distinguishing between shocks with permanent and transitory effects is perhaps done more naturally in models which allow for integrated variables. They will be discussed later.
5. Sign restrictions may be imposed on the impulse responses (for example, Canova and De Nicoló 2003; Uhlig 2005), that is, one may want to require that certain shocks have positive or negative effects on certain variables. For example, a restrictive monetary shock should reduce the inflation rate.

Integrated and Cointegrated Variables

If the VAR operator has unit roots, that is, $\det A(z) = 0$ for $z = 1$, then the variables have stochastic trends. Variables with such trends are called integrated. They can be made stationary by differencing. Moreover, they are called cointegrated if stationary linear combinations exist. If the VAR model contains integrated and cointegrated variables, impulse response analysis can still be performed as for stationary processes. For the latter processes the Φ_i's go to zero for $i \to m$ and, hence, the marginal response to an impulse to a stationary process is transitory, that is, the effect goes to zero as time goes by. In contrast, some impulses have permanent effects in cointegrated systems. In fact, in a K-dimensional system with $r < K$ cointegration relations, at least $K - r$ of the K shocks have permanent effects and at most r shocks have transitory effects (King et al. 1991; Lütkepohl 2005, ch. 9). These facts open up the possibility to find identifying restrictions for the structural innovations by taking into account the cointegration properties of the system.

Estimation of Impulse Responses

Estimation of reduced form and structural form parameters of VAR processes is usually done by least squares, maximum likelihood or Bayesian methods. Estimates of the impulse responses are then obtained from the VAR parameter estimates. Suppose the VAR coefficients are contained in a vector $\boldsymbol{\alpha}$ and denote its estimator by $\hat{\boldsymbol{\alpha}}$. Any specific impulse response coefficient θ is a (nonlinear) function of $\boldsymbol{\alpha}$ and may be estimated as $\widehat{\theta} = \theta(\hat{\boldsymbol{\alpha}})$. If $\hat{\boldsymbol{\alpha}}$ is asymptotically normal, that is, $\sqrt{T}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\hat{\boldsymbol{\alpha}}})$, then, under general conditions, $\hat{\theta}$ is also asymptotically normally distributed, $\sqrt{T}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\theta^2\right)$. The variance of the asymptotic distribution is $\sigma_\theta^2 = \frac{\partial \theta}{\partial \boldsymbol{\alpha}'} \Sigma_{\hat{\boldsymbol{\alpha}}} \frac{\partial \theta}{\partial \boldsymbol{\alpha}}$. Here $\partial\theta/\partial\boldsymbol{\alpha}$ denotes the vector of first order partial derivatives of θ with respect to the elements of $\boldsymbol{\alpha}$ (see Lütkepohl 1990, for the precise expressions). This result can be used for setting up asymptotic confidence intervals for impulse responses in the usual way.

Asymptotic normality of $\widehat{\theta}$ requires that σ_θ^2 is non-zero, which follows if $\Sigma_{\hat{\boldsymbol{\alpha}}}$ is non-singular and $\partial\theta/\partial\alpha \neq 0$. In general the covariance matrix $\Sigma_{\hat{\boldsymbol{\alpha}}}$ will not be nonsingular for cointegrated systems, for example. Moreover, the impulse responses generally consist of sums of products of the VAR coefficients and, therefore, the partial derivatives will also be sums of products of such coefficients. Consequently, the partial derivatives will also usually be zero in parts of the parameter space. Thus, $\sigma_\theta^2 = 0$ may hold and, hence, $\hat{\theta}$ may actually converge at a faster rate than $\sqrt{T}$ in parts of the parameter space (cf. Benkwitz et al. 2000).

Even under ideal conditions where the asymptotic theory holds, it may not provide a good guide for small sample inference. Therefore, bootstrap

methods are often used to construct confidence intervals for impulse responses (for example, Kilian 1998; Benkwitz et al. 2001). If one uses these methods, deriving explicit forms of the analytical expressions for the asymptotic variances of the impulse response coefficients can be avoided. Unfortunately, bootstrap methods generally do not overcome the problems due to zero variances in the asymptotic distributions of the impulse responses. In fact, they may provide confidence intervals which do not have the desired coverage level even asymptotically (Benkwitz et al. 2000).

Confidence bands for impulse response functions can also be constructed with Bayesian methods (for example, Koop 1992). Prior information on the VAR parameters or the impulse responses can in that case be considered. It is not uncommon to report confidence intervals for individual impulse response coefficients and connecting them to get a confidence band around an impulse response function. This approach has been criticized by Sims and Zha (1999), who propose likelihood- characterizing error bands instead.

Extensions

There are a number of extensions to the models and impulse response functions considered so far. For example, all observed variables are treated as endogenous. A main criticism regarding problematic exogeneity assumptions in classical simultaneous equations models is thereby accounted for. On the other hand, this approach often results in heavily parameterized models and imprecise estimates. Therefore, it is occasionally desirable to classify some of the variables as exogenous or consider partial models where we condition on some of the variables which remain unmodelled. In this case one may be interested in tracing the effects of changes in the exogenous or unmodelled variables on the endogenous variables. The resulting impulse response functions are often referred to as dynamic multipliers in the literature on simultaneous equations (see Lütkepohl 2005, for an introductory treatment). The inference problems related to these quantities are similar to those discussed earlier for VAR impulse responses.

It was also acknowledged in the related literature that finite order VAR models are at best good approximations to the actual data generation processes of multiple time series. Therefore, inference for impulse responses was also considered under the assumption that finite order VAR processes are fitted to data generated by infinite order processes (for example, Lütkepohl 1988; Lütkepohl and Saikkonen 1997).

Impulse responses associated with linear VAR models have the property of being time invariant and their shape is invariant to the size and direction of the impulses. These features make it easy to represent the reactions of the variables to impulses hitting the system in a small set of graphs. Such responses are often regarded as unrealistic in practice, where, for instance, a positive shock may have a different effect from a negative shock or the effect of a shock may depend on the state of the system at the time when it is hit. Hence, the linear VAR models are too restrictive for some analyses. These problems can be resolved by considering nonlinear models. Although nonlinear models have their attractive features for describing economic systems or phenomena, their greater flexibility makes them more difficult to interpret properly. In fact, it is not obvious how to define impulse responses of nonlinear models in a meaningful manner. Gallant et al. (1993) proposed so-called *conditional moment profiles* which may give useful information on important features of nonlinear multiple time series models. For example, one may consider quantities of the general form $E[g(y_{t+h})|y_t + \xi, \Omega_{t-1}] - E[g\{y_{t+h}\})|y_t, \Omega_{t-1}]$, $h = 1,2,\ldots$, where $g(\cdot)$ denotes some function of interest, ξ represents the impulses hitting the system at time t, and $\Omega_{t-1} = (y_{t-1}, y_{t-2},\ldots)$ denotes the history of the variables at time t. In other words, the conditional expectation of some quantity of interest, given the history of y_t in period t, is compared to the conditional expectation that is obtained if a shock ξ occurs at time t. For example, defining $g\{y_{t+h}) = [y_{t+h} - E\{y_{t+h}|\Omega_{t+h-1})][y_{t+h} - E\{y_{t+h}|\Omega_{t+h-1})]'$ results in conditional volatility profiles, which may be compared to a baseline

profile obtained for a specific history of the process and a zero impulse. Clearly, in general the conditional moment profiles depend on the history Ω_{t-1} as well as the impulse ξ. Similar quantities were also considered by Koop et al. (1996), who called them *generalized impulse responses* (see also Pesaran and Shin 1998).

Although these quantities may be interesting to look at, they depend on t, h, and ξ. Hence, there is a separate impulse response function for each given t and ξ. In empirical work it will therefore be necessary to summarize the wealth of information in the conditional moment profiles in a meaningful way – for instance, by considering summary statistics. In practice, an additional obstacle is that the actual data generation process is unknown and estimated models are available at best. In that case, the conditional moment profiles or generalized impulse responses will be estimates, and it would be useful to have measures for their sampling variability. It is not clear how this additional information may be computed and presented in the best way in practice.

See Also

- ▶ Cointegration
- ▶ Long Run and Short Run
- ▶ Measurement Error Models
- ▶ Multiplier Analysis
- ▶ Structural Vector Autoregressions
- ▶ Vector Autoregressions

Bibliography

Amisano, G., and C. Giannini. 1997. *Topics in structural VAR econometrics*. 2nd ed. Berlin: Springer.

Benkwitz, A., H. Lütkepohl, and M. Neumann. 2000. Problems related to bootstrapping impulse responses of autoregressive processes. *Econometric Reviews* 19: 69–103.

Benkwitz, A., H. Lütkepohl, and J. Wolters. 2001. Comparison of bootstrap confidence intervals for impulse responses of German monetary systems. *Macroeconomic Dynamics* 5: 81–100.

Blanchard, O., and D. Quah. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79: 655–673.

Canova, F., and G. De Nicoló. 2003. On the sources of business cycles in the G-7. *Journal of International Economics* 59: 77–100.

Christiano, L., M. Eichenbaum, and C. Evans. 1996. The effects of monetary policy shocks: Evidence from the flow of funds. *The Review of Economics and Statistics* 78: 16–34.

Gallant, A., P. Rossi, and G. Tauchen. 1993. Nonlinear dynamic structures. *Econometrica* 61: 871–907.

Kilian, L. 1998. Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics* 80: 218–230.

King, R., C. Plosser, J. Stock, and M. Watson. 1991. Stochastic trends and economic fluctuations. *American Economic Review* 81: 819–840.

Koop, G. 1992. Aggregate shocks and macroeconomic fluctuations: A Bayesian approach. *Journal of Applied Econometrics* 7: 395–411.

Koop, G., M. Pesaran, and S. Potter. 1996. Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics* 74: 119–147.

Lütkepohl, H. 1988. Asymptotic distribution of the moving average coefficients of an estimated vector autoregressive process. *Econometric Theory* 4: 77–85.

Lütkepohl, H. 1990. Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *The Review of Economics and Statistics* 72: 116–125.

Lütkepohl, H. 2005. *New introduction to multiple time series analysis*. Berlin: Springer.

Lütkepohl, H., and P. Saikkonen. 1997. Impulse response analysis in infinite order cointegrated vector autoregressive processes. *Journal of Econometrics* 81: 127–157.

Pesaran, M., and Y. Shin. 1998. Generalized impulse response analysis in linear multivariate models. *Economics Letters* 58: 17–29.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.

Sims, C., and T. Zha. 1999. Error bands for impulse responses. *Econometrica* 67: 1113–1155.

Uhlig, H. 2005. What are the effects of monetary policy on output? Results from an agnostic identification procedure. *Journal of Monetary Economics* 52: 381–419.

Imputation

Murray N. Rothbard

Keywords

Aristotle; Austrian School; Böhm-Bawerk, E. von; Factor prices; Imputation; Marginal productivity theory; Menger, C.; Socialist calculation debate; Subjective theory of value;

Wieser, F. F. von; Mises, L. E. von; Hayek, F. A. von

JEL Classifications
D2

'Imputation' is a term introduced into economics as *Zurechnung* by the Austrian School economist Friedrich Freiherr von Wieser (1889). The term was a legal one, and the analogy was based on the legal method by which the jurist *imputes* guilt or liability to one or another criminal or tortfeasor. Imputation was a central concern of the Austrian School, since its analysis centred on the nature of the means–ends relationship (Mises 1949) and on the process by which the subjective valuations and value-preferences of individual consumers 'impute' value to the goods being produced. As Carl Menger, founder of the Austrian School, pointed out, the valuations by consumers of their satisfactions, or ends, impute values to the consumer goods, the means, that are expected to satisfy those wants (Menger 1871). And since producers' goods are only means to the production and sale of consumer goods, the values of the factors of production will in turn be determined by and be equal to the expected values of the consumer goods to the consumers. In short, values are 'imputed' back to the prices of the factors of production; the rents of Champagne land are high because the consumers value the champagne highly, and not the other way round. 'Costs' of resources are reflections of the value of products forgone.

While this process was clear in principle, there were considerable difficulties in working out the specifics. Essentially, Menger and his student Böhm-Bawerk stuck close to the realities of the market process, and focused on value imputation as a process of estimating how much of a product would be lost if the producer were deprived of one unit of a factor. Wieser, on the other hand, presumed that the marginal value of each factor could be found with great precision; in doing so, he assumed illegitimately that subjective values can be added and multiplied to arrive at the total value of a quantity of goods. But by its nature subjective value is an expression of ordinal preferences and therefore can neither be added nor measured.

The modern theory of marginal productivity has essentially solved these problems and shown how values of products can be imputed back to productive factors. One exception is the current assumption that the existence of variable proportions solves the problem of pricing factors and leaves no theoretical room for arbitrary bargaining between factor owners. But the more important solution depends on whether factors are purely specific to one line of production or are relatively non-specific, that is, can be employed in the production of more than one good. If two factors are each purely specific to a given product, then, even if their proportions are variable, there is still no principle by which the market can determine their relative prices except by arbitrary bargaining (Mises 1949, p. 336). In the real world, of course, the existence of such purely specific factors, and hence the scope for such bargaining, will be extremely limited.

The other important point is that values cannot be added or divided, and that the imputation process takes place, not automatically or precisely in an abstract realm of 'values', but only concretely and by trial and error, in the realistic market process of changing prices. In other words, although consumers can evaluate consumer goods and determine their prices directly by valuation, the prices of productive factors are only determined indirectly through market prices and entrepreneurial trial and error. There is no direct, abstract or pure process of imputing values.

This problem became strikingly relevant during the well-known debate over the Mises–Hayek demonstration that socialist governments cannot calculate economically. Joseph Schumpeter brusquely dismissed this contention with the statement that economic calculation under socialism follows 'from the elementary proposition that consumers in evaluating ("demanding") consumers' goods *ipso facto* also evaluate the means of production which enter into the production of these goods' (Schumpeter 1942, p. 175). Hayek's perceptive reply points out that the *'ipso facto'* assumes complete knowledge of values, demands, scarcities, and so on, to be 'given' to everyone,

thereby ignoring the reality of the universal lack of complete knowledge, as well as the necessary function of the market economy, and the market price system, in conveying knowledge to all its participants (Hayek 1945).

The analysis of imputation began in a neglected work of Aristotle, the *Topics.* Here, Aristotle analysed the ends–means relationship, and pointed out that the means, or 'instruments of production', necessarily derive their value from the ends, the final products useful to man, 'the instruments of action'. The more desirable the final good, the more valuable will be the means to arrive at the product. Aristotle introduced the theme of marginality by stating that, if the addition of a good *A* to an already desirable good *C* yields a more desirable result than the addition of good *B,* then *A* will be more highly valued than B. Indeed, he also added a pre-Böhm- Bawerkian note by stressing the differential value of the loss rather than the addition of a good. Good *A* will be more valuable than *B* if the loss of *A* is considered to be worse than the loss of B. While critics have noted that Aristotle only slightly applied his analysis to the economic realm, his imputation theory was still an important contribution to the general theory of action of which economic theory is a highly developed part (Spengler 1955).

See Also

- ▶ Austrian Economics
- ▶ Marginal Productivity Theory

Bibliography

Aristotle 1928. *Topica.* Trans. W.A. Pickard-Cambridge and included. In *The works of Aristotle,* vol. 1, ed. W.D. Ross. Oxford: Clarendon Press.

Hayek, F.A. 1926. Some remarks on the problem of imputation. In *Money, capital, and fluctuations: Early essays*, ed. F.A. Hayek. Chicago: University of Chicago Press. 1984.

Hayek, F.A. 1945. The use of knowledge in society. In *Individualism and economic order*, ed. F.A. Hayek. Chicago: University of Chicago Press. 1948.

Kauder, E. 1965. *A history of marginal utility theory.* Princeton: Princeton University Press.

Menger, C. 1871. *Principles of economics*. Glencoe: Free Press, 1950.

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

Spengler, J. 1955. Aristotle on economic imputation and related matters. *Southern Economic Journal* 21: 371–389.

Stigler, G. 1941. *Production and distribution theories: The formative period*. New York: Macmillan.

von Wieser, F. 1889. *Natural value*, 1956. New York: Kelley & Millman.

von Mises, L. 1949. *Human action: A treatise on economics,* 3rd ed. Chicago: Regnery, 1966.

Incentive Compatibility

John O. Ledyard

I

Abstract

Incentive compatibility – a characteristic of mechanisms whereby each agent knows that his best strategy is to follow the rules, no matter what the other agents will do – is desirable because it promotes the achievement of group goals. But it is elusive because pervasive opportunities exist for misbehaviour, such as by misrepresenting preferences. This article reviews attempts to solve or at least to manage the incentive compatibility problem. Incentive compatibility provides a basic constraint on the possibilities for normative analysis, and so serves as the fundamental interface between what is desirable and what is possible in a theory of organizations.

Keywords

Allocation mechanisms; Auctions; Bayes' equilibrium; Borda, J.-C. de; Capital budgeting; Central planning; Cobb–Douglas functions; False preferences; Free rider problem; Games of incomplete information; Incentive compatibility; Ledyard, J. O.; Majority rule; Market failure; Mechanism design; Monotonicity; Nash equilibrium; No-trade option; Offer curves; Pareto efficiency; Principal and agent; Public enterprise management;

Public goods; Regulation of monopoly; Revelation principle; Self-selection; Social welfare functions; Synthetic markets; Transfer pricing; Tâtonnement processes; von Neumann–Morgenstern utility function

JEL Classifications
D0

Allocation mechanisms, organizations, voting procedures, regulatory bodies, and many other institutions are designed to accomplish certain ends such as the Pareto- efficient allocation of resources or the equitable resolution of disputes. In many situations it is relatively easy to conceive of feasible processes; processes which will accomplish the goals if all participants follow the rules and are capable of handling the informational requirements. Examples of such mechanisms include marginal cost pricing, designed to attain efficiency, and equal division, designed to attain equity. Of course once a feasible mechanism is found, the important question then becomes whether such a mechanism is also informationally feasible and compatible with 'natural' incentives of the participants. Incentive compatibility is the concept introduced by Hurwicz (1972, p. 320) to characterize those mechanisms for which participants in the process would not find it advantageous to violate the rules of the process.

The historical roots of the idea of incentive compatibility are many and deep. As was pointed out in one of a number of recent surveys,

> the concept of incentive compatibility may be traced to the 'invisible hand' of Adam Smith who claimed that in following individual self-interest the interests of society might be served. Related issues were a central concern in the 'Socialist Controversy' which arose over the viability of a decentralized socialist society. It was argued by some that such societies would have to rely on individuals to follow the rules of the system. Some believed this reliance was naive; others did not. (Groves and Ledyard 1986, p. 1)

Further, the same issues have arisen in the design of voting procedures. Concepts and problems related to incentives were already identified and documented in the 18th century in discussions of proposals by Borda to provide alternatives to majority rule committee decisions. (See ▶ Strategy-proof Allocation Mechanisms for further information on voting procedures.)

Incentive compatibility is both desirable and elusive. The desirability of incentive compatibility can be easily illustrated by considering public goods, goods such that one consumer's consumption of them does not detract from another consumer's simultaneous consumption of that good. The existence of these collective consumption commodities creates a classic situation of *market failure*; the inability of markets to arrive at a Pareto-optimal allocation. It was commonly believed, prior to Groves and Ledyard (1977), that in economies with public goods it would be impossible to devise a decentralized process that would allocate resources efficiently since agents would have an incentive to 'free ride' on others' provision of those goods in order to reduce their own share of providing them. Of course Lindahl (1919) had proposed a feasible process which mimicked markets by creating a separate price for each individual's consumption of the public good. This designed process was, however, rejected as unrealistic by those who recognized that these 'synthetic markets' would be shallow (essentially monopsonistic) and therefore buyers would have no incentive to treat prices as fixed and invariant to their demands. The classic quotation is '... it is in the selfish interest of each person to give *false* signals, to pretend to have less interest in a given collective consumption activity than he really has...' (Samuelson 1954, pp. 388–9). Allocating public goods efficiently through Lindahl pricing would be feasible and successful if consumers followed the rules; but, it would not be successful since the mechanism is not incentive compatible. If buyers do not follow the rules, efficient resource allocation will not be achieved and the goals of the design will be subverted because of the motivations of the participants. Any institution or rule, designed to accomplish group goals, must be incentive compatible if it is to perform as desired.

The elusiveness of incentive compatibility can be most easily illustrated by considering a situation with only private goods. Economists generally model behaviour in private goods markets by assuming that buyers and sellers 'follow the rules' and take prices as given. It is now known, however, that as long as the number of agents is finite then any one of them can still gain by misbehaving and, furthermore, can do so in a way which can not be detected by anyone else. The explanation is provided in two steps. First, if there are a finite number of traders, and none have a perfectly elastic offer curve (which will be true if preferences are non- linear) then one trader can gain by being able to control prices. For example, a buyer would want to set price where his marginal benefit equalled his marginal outlay and thereby gain monopsonistic benefits. Of course, if the others know that buyer's demand curve (either directly or through inferences based on revealed preference) then they would know that the buyer was not 'taking prices as given' and could respond with a suitable punishment against him. This brings us to our second step. Even though others can monitor and prohibit price setting behaviour, our benefit- seeking monopsonist has another strategy which can circumvent this supervision. He calculates a (false) demand curve which, when added to the others' offer curves, produces an equilibrium price equal to that which he would have set if he had direct control. He then calculates a set of preferences which yields that demand curve and participates in the process *as if he had these (false) preferences*. Usually this involves simply acting as if one has a slightly lower demand curve than one really does. Since preferences are not able to be observed by others, he can follow this behaviour which looks like it is price-taking, and therefore 'legal', and can do individually better. The unfortunate implication of such concealed misbehaviour is that the mechanism performs other than as intended. In this case, resources are artificially limited and too little is traded to attain efficiency.

In 1972 Hurwicz established the validity of the above intuition. His theorem can be precisely stated after the introduction of some notation and a framework for further discussion.

The Impossibility Theorem

The key concepts include economic environments, allocation mechanisms, incentive compatibility, the no-trade option, and Pareto-efficiency. We take up each in turn.

An *economic environment*, those features of an economy which are to be taken as given throughout the analysis, includes a description of the agents, the feasible allocations they have available and their preferences for those allocations. While many variations are possible, I concentrate here on a simple model. Agents (consumers, producers, politicians, etc.) are indexed by $i = 1, \ldots, n$. X is the set of feasible allocations where $x = (x^i, \ldots, x^n)$ is a typical element of X. (An exchange environment is one in which X is the set of all $x = (x^1, \ldots, x^n)$ such that $x^i \geq 0$ and $\sum x^i = \sum w^i$, where w^i is i's initial endowment of commodities.) Each agent has a selfish utility function $u^i(x^i)$. The environment is $e = [I, X, u^1, \ldots, u^n]$. A crucial fact is that initially *information is dispersed* since i, and only i, knows u^i. We identify the specific knowledge i initially has as i's *characteristic,* e^i. In our model, $e^i = u^i$.

Although there are many variations in models of allocation mechanisms, I begin with the one introduced by Hurwicz (1960). An *allocation mechanism* requests information from the agents and then computes a feasible allocation. It requests information in the form of messages m^i from agent i through a *response* function $f^i(m^i, \ldots, m^n)$. Agent i is told to report $f^i(m, e^i)$ if others have reported m and i's characteristic is e^i. An equilibrium of these response rules, for the environment e, is a joint message m such that $m^i = f^i(m, e^i)$ for all i. Let $\mu(e, f)$ be the set of equilibrium messages for the response functions f in the environment e. The allocation mechanism computes a feasible allocation x by using an *outcome* function $g(m)$ on equilibrium messages. The net result of all of this in the environment e is the allocation $g[\mu(e, f)] = x$ *if all i follow the rules*, f. Thus, for example, the *competitive mechanism* requests agents to send their demands as a function of prices which are in turn computed on the basis of the aggregate demands reported by the consumers. In equilibrium, each agent is simply allocated their stated demand. (An alternative mechanism, yielding exactly the

same allocation in one iteration, would request the demand function and then compute the equilibrium price and allocation for the reported demand functions.) It is well known, for exchange economies with only private goods, that if agents report their true demands then the allocations computed by the competitive mechanism will be Pareto-optimal.

It is obviously important to be able to identify those mechanisms, those rules of communication, that have the property that they are self-enforcing. We do that by focusing on a class of mechanisms in which each agent gains nothing, and perhaps even loses, by misbehaving. While a multitude of misbehaviours could be considered it is sufficient for our purposes to consider a slightly restricted range. In particular we can concentrate on undetectable behaviour, behaviour which no outside agent can distinguish from that prescribed by the mechanism. We model this limitation on behaviour by requiring the agent to restrict his misrepresentations to those which are consistent with some characteristic he might have. An allocation mechanism is said to be *incentive compatible* for all environments in the class E if there is no agent i and no environment e in E and no characteristic e^{*i} such that (e/e^{*i}) is in E (where (e/e^{*i}) is the environment derived from e by replacing e^i with e^{*i}) and such that

$$u^i\{g[\mu(e,f)],e^i\} < u^i\{g[\mu(e/e^{*i},f)],e^i\}$$

where $u^i(x^*, e^i)$ is i's utility function in the environment e. That is, no agent can manipulate the mechanism by pretending to have a characteristic different from the true one and do better than acting according to the truth. The agent has an incentive to follow the rules and the rules are compatible with his motivations.

Incentive compatibility is at the foundation of the modern *theory of implementation*. In that theory, one tries to identify conditions under which a particular social choice rule or performance standard, $P : E \to X$, can be recreated by an allocation mechanism under the hypothesis that individuals will follow their self-interest when they participate in the implementation process. In our language, the rule P is implementable if and only if there is an incentive compatible mechanism (f, g) such that $g[\mu(e, f)] = P(e)$ for all e in E. The theory of implementation seeks to answer the question 'which P are implementable?' We will see some of the answers below for P which select from the set of Pareto-efficient allocations. Those interested in more general goals and performance standards should consult Dasgupta et al. (1979) or Postlewaite and Schmeidler (1986).

An allocation mechanism is said to have the *no trade-option* if there is an allocation θ at which each participant may remain. In exchange environments the initial endowment is usually such an allocation. Mechanisms with a no-trade option are non-coercive in a limited sense. If an allocation mechanism possesses the no- trade option then the allocation it computes for an environment e, if agents follow the rules, must leave everyone at least as well off, using the utility functions for e, as they are at θ. That is, for all i and all e in E

$$u^i\{g[\mu(e,f)],e^i\} > u^i(\theta,e^i).$$

An allocation mechanism is said to be *Pareto-efficient in E* if the allocations selected by the mechanism, when agents follow the rules, are Pareto-optimal in e. That is, for each e in E, there is no allocation x^* in X such that, for all i,

$$u^i(x^*,e^i) \geq u^i\{g[\mu(e,f)],e^i\}$$

with strict inequality for some i.

With this language and notation, Hurwicz's theorem on the elusive nature of incentive compatibility in private markets, subsequently expanded by Ledyard and Roberts (1974) to include public goods environments, can now be easily stated. *Theorem*: In classical (public or private) economic environments with a finite number of agents, there is no incentive compatible allocation mechanism which possesses the no-trade option and is Pareto-efficient. (Classical environments include pure exchange environments with Cobb–Douglas utility functions.)

A more general version of this theorem, in the context of social choice theory, has been proven by Gibbard (1973) and Satterthwaite (1975) with

the concept of a 'non-dictatorial social choice function' replacing that of a 'mechanism with the no- trade option'. (See ▶ Strategy-proof Allocation Mechanisms.)

There are a variety of possible reactions to this theorem. One is simply to give up the search for solutions to market failure since the theorem seems to imply that one should not waste any effort trying to create institutions to allocate resources efficiently. A second is to notice that, at least in private markets, if there are a very large number of individuals in each market then efficiency is 'almost' attainable (see Roberts and Postlewaite 1976). A third is to recognize that the behaviour of individuals will generally be different from that implicitly assumed in the definition of incentive compatibility. A fourth is to accept the inevitable, lower one's sights, and look for the 'most efficient' mechanism among those which are incentive compatible and satisfy a voluntary participation constraint. We consider the last two options in more detail.

Other Behaviour: Nash Equilibrium

If a mechanism is incentive compatible, then each agent knows that his best strategy is to follow the rules according to his true characteristic, *no matter what the other agents will do*. Such a strategic structure is referred to as a dominant strategy game and has the property that no agent need know or predict anything about the others' behaviour. In mechanisms which are not incentive compatible, each agent must predict what others are going to do in order to decide what is best. In this situation agents' behaviour will not be as assumed in the definition of incentive compatibility. What it will be continues to be an active research topic and many models have been proposed. Since most of these are covered in Groves and Ledyard (1986), I will concentrate on the two which seem most sensible. Both rely on game-theoretic analyses of the strategic possibilities. The first concentrates on the outcome rule, g, and postulates that agents will not choose messages to follow the specifications of the response functions but to do the best they can against the messages sent by others. Implicitly this assumes that there is some type of iterative process (embodied in the response rules) which allows revision of one's message in light of the responses of others. We can formalize this presumed strategic behaviour in a new concept of incentive compatibility. An allocation mechanism (f, g) is called *Nash incentive compatible* for all environments in E if there is no environment e, no agent i, and no message m^{*i} which i can send such that

$$u^i\left(g\left[\mu(e,f)/m^{*i}, e^i\right]\right) > u^i\left(g\left[\mu(e,f), e^i\right]\right)$$

where $\mu(e, f)$ is the 'equilibrium' message of the response rules f in the environment e, $g(m)$ is the outcome rule, and $[m/m^{*i}]$ is the vector m where m^{*i} replaces m^i. In effect this requires the equilibrium messages of the response rules to be Nash equilibria in the game in which messages are strategies and payoffs are given by $u[g(m)]$. It was shown in a sequence of papers written in the late 1970s, including those by Groves and Ledyard (1977), Hurwicz (1979), Schmeidler (1980), and Walker (1981), that Nash incentive compatibility is not elusive. The effective output of that work was to establish the following. *Theorem*: In classical (public or private) economic environments with a finite number of agents, there are many Nash incentive compatible mechanisms which possess the no-trade option and are Pareto-efficient.

With a change in the predicted behaviour of the participants in the mechanism, in recognition of the fact that in the absence of dominant strategies agents must follow some other self-interested strategies, the pessimism of the Hurwicz theorem is replaced by the optimistic prediction of a plethora of possibilities. (See Dasgupta et al. (1979), Postlewaite and Schmeidler (1986) and Groves and Ledyard (1986) for comprehensive surveys of these results including many for more general social choice environments.) Although it remains an unsettled empirical question whether participants will indeed behave this way, there is a growing body of experimental evidence that seems to me to support the behavioural hypotheses

underpinning Nash incentive compatibility, especially in iterative tâtonnement processes.

Other Behaviour: Bayes' Equilibrium

The second approach to modelling strategic behaviour of agents in mechanisms, when dominant strategies are not available, is based on Bayesian decision theory. These models, called *games of incomplete information* (see Myerson 1985), concentrate on the beliefs of the players about the situation in which they find themselves. In the simplest form, it is postulated that there is a common knowledge (everyone knows that everyone knows that...) probability function, $\pi(e)$, which describes everyone's prior beliefs. Each agent is then assumed to choose that message which is best against the expected behaviour of the other agents. The expected behaviour of the other agents is also constrained to be 'rational' in the sense that it should be best against the behaviour of others. This presumed strategic behaviour is embodied in a third type of incentive compatibility. (It could be argued that the concept of incentive compatibility remains the same, based on non- cooperative behaviour in the game induced by the mechanism, while only the presumed information structure and sequence of moves required to implement the allocation mechanism are changed. Such a view is not inconsistent with that which follows.) An allocation mechanism (f, g) is called *Bayes incentive compatible* for all environments in E given π on E if there is no environment e^*, no agent i, and no message m^{*i} which i can send such that

$$\int u^i\{g[\mu(e,f)/m^{*i}], e^{*i}\}\mathrm{d}\pi(e|e^{*i}\} > \int u^i\{g[\mu(e,f), e^{*i}]\mathrm{d}\pi(e,|e^{*i})\}$$

where, as before, μ is the equilibrium message vector and g is the outcome rule. Further, $\pi(e/e^{*i})$ *is the conditional probability measure on e* given e^{*i}, and u^i is a von Neumann–Morgenstern utility function. In effect, this requires the equilibrium messages of the response rules to be Bayes equilibrium outcomes of the incomplete information game with messages as strategies, payoffs $u[g(m)]$ and common knowledge prior π.

There are two types of results which deal with the possibilities for Bayes incentive compatible design of allocation mechanisms, neither of which is particularly encouraging. The first type deals with the possibilities for incentive compatible design which is independent of the beliefs. The typical theorem is illustrated by the following result proven by Ledyard (1978). *Theorem*: In classical economic environments with a finite number of agents, there is no Bayes incentive compatible mechanism which possesses the no-trade option and is Pareto-efficient *for all* π *on* E. Understanding this result is easy when one realizes that any mechanism (f, g) is Bayes incentive compatible for all π for all e in E if and only if it is (Hurwicz) incentive compatible for all e in E. Thus the Hurwicz impossibility theorem again applies.

The second type of result is directed towards the possibilities for a specific prior π; that is, towards what can be done if the mechanism can depend on the common knowledge beliefs. The most general characterizations of the possibilities for Bayes incentive compatible design can be found in Palfrey and Srivastava (1987) and Postlewaite and Schmeidler (1986). They have shown that two conditions, called monotonicity and self-selection, are necessary and sufficient for a social choice correspondence to be implementable in the sense that there is a Bayes incentive compatible mechanism that reproduces that correspondence. The details of these conditions are not important. What is important is that many correspondences do not satisfy them. In particular, there appear to be many priors π and many sets of environments E for which there is no mechanism which is Bayes incentive compatible, provides a no-trade option and is Pareto-efficient. Thus, impossibility still usually occurs even if one allows the mechanism to depend on the prior.

One recent avenue of research which promises some optimistic counterweight to these negative results can be found in Palfrey and Srivastava (1987). In much the same way that the natural move from Hurwicz incentive compatibility to Nash incentive compatibility created

opportunities for incentive compatible design, these authors have shown that a move back towards dominant strategies may also open up possibilities. Refinements arise by varying the equilibrium concept in a way that reduces the number of (Bayes or Nash) equilibria for a given e or π. Moore and Repullo use subgame perfect Nash equilibria. Palfrey and Srivastava eliminate weakly dominated strategies from the set of Nash equilibria. They have discovered that, in pure exchange environments, virtually all performance correspondences are implementable if behaviour satisfies these refinements. In particular, any selection from the Pareto-correspondence is implementable for these refinements, and so there are many refined-Nash incentive compatible mechanisms which are Pareto-efficient and allow a no-trade option. It is believed that these results will transfer naturally to refinements of Bayes equilibria, but the research remains to be done.

Incentive Compatibility as a Constraint

Another of the reactions to the Hurwicz impossibility result is to accept the inevitable, to view incentive compatibility as a constraint, and to design mechanisms to attain the best level of efficiency one can. If full efficiency is possible, it will occur as the solution. If not, then one will at least find the second-best allocation mechanism. Examples of this rapidly expanding research literature include work on optimal auctions (Harris and Raviv 1981; Matthews 1983; Myerson 1981), the design of optimal contracts for the principle-agent problem, and the theory of optimal regulation (Baron and Myerson 1982). As originally posed by Hurwicz (1972, pp. 299–301), the idea is to adopt a social welfare function $W(x, e)$, a measure of the social welfare attained from the allocation x if the environment is e and then to choose the mechanism (f, g) to maximize the (expected) value of W subject to the 'incentive compatibility constraints', the constraint that the rules (f, g) be consistent with the motivations of the participants. One chooses (f, g) to

$$\text{maximize} \int W\{g[\mu(e,f)], e\}\mathrm{d}\pi(e)$$

subject to, for every i, every e, and every e^{*i},

$$\int u^i\{g[\mu(e/e^{*i},f)], e^i\}\mathrm{d}\pi(e|\,e^i) \leq \int u^i(g[\mu(e,f), e^i])\mathrm{d}\pi(e|\,e^i).$$

As formalized here the incentive compatibility constraints embody the concept of Bayes incentive compatibility. Of course, other behavioural models could be substituted as appropriate.

Sometimes a voluntary participation constraint, related to the no-trade option of Hurwicz, is added to the optimal design problem. One form of this constraint requires that (f, g) also satisfy, for every i and every e,

$$\int u^i\{g[\mu(e)], e^i\}\mathrm{d}\pi(e|\,e^i) \geq \int u^i(\theta[e], e^i)\mathrm{d}\pi(e|\,e^i).$$

In practice this optimization can be a difficult problem since there are a large number of possible mechanisms (f, g). However, an insight due to Gibbard (1973) can be employed to reduce the range of alternatives and simplify the analysis. Now called the *revelation principle*, the observation he made was that, to find the maximum, it is sufficient to consider only mechanisms, called direct revelation mechanisms, in which agents are asked to report their own characteristics. The reason is easy to see. Suppose that (f^*, g^*) solves the maximum problem. Let (F^*, G^*) be a new (direct revelation) mechanism defined by $F^{*i}(m, e^i) = e^i$ and $G^*(m) = g[\mu(m, f)]$. Each i is told to report his characteristic and then G^* computes the allocation by computing that which would have been chosen if the original mechanism (f, G^*) had been used honestly in the reported environment. (F^*, G^*) yields the same allocation as (f^*, g^*), *if each agent reports the truth*. But the incentive compatibility constraints, which (f^*, g^*) satisfied, ensure that each agent will want to report truthfully. Thus, whatever can be done, by any

arbitrary mechanism subject to the Bayes incentive compatibility constraints, can be done with direct revelation mechanisms subject to the constraint that each agent wants to report their true characteristic. One need only choose a function $G : E \to X$ to

$$maximize \int W\{G(e), e\} d\pi(e)$$

subject to, for every *i, e* and e^i,

$$\int u^i\{G(e/e^{*i}), e^i\} \mathrm{d}\pi(e|\, e^i) \leq \int u^i[G(e), e^i] \mathrm{d}\pi(e|\, e^i),$$

and

$$\int u^i[G(e), e^i] \mathrm{d}\pi(e|\, e) \geq \int u^i(\theta[e], e^i) \mathrm{d}\pi(e|\, e^i).$$

There are at least two problems with this approach to organizational design. The first is that the choice of mechanism depends crucially on the prior beliefs, π. This is a direct result of the use of Bayes incentive compatibility in the constraints. Since the debate is still open let me simply summarize some of the arguments. One is that if the mechanism chosen for a given situation does not depend on common knowledge beliefs then we would not be using all the information at our disposal to pursue the desired goals and would do less than is possible. Further, since the beliefs are common knowledge we can all agree as to their validity (misrepresentation is not an issue) and therefore to their legitimate inclusion in the calculations. An argument is made against this on the practical grounds that one need only consider actual situations, such as the introduction of new technology by a regulated utility or the acquisition of a major new weapons system by the government, to understand the difficulties involved in arriving at agreements about the particulars of common knowledge. Another argument against is based on the feeling that mechanisms should be robust. A 'good' mechanism should be able to be described in terms of its mechanics and, while it probably should have the capacity to incorporate the common knowledge relevant to the current situation, it should be capable of being used in many situations. How to capture these criteria in the constraints or the objective function of the designer remains an open research question.

The second problem with the optimal auction approach to organizational design is the reliance on the revelation principle. Restricting attention to direct revelation mechanisms, in which an agent reports his entire characteristic, is an efficient way to prove theorems, but it provides little guidance for those interested in actual organization design. For example it completely ignores the informational requirements of the process and any limitations, if any, in the information processing capabilities of the agents or the mechanism. Writing down one's preferences for all possible consumption patterns is probably harder than writing down one's entire demand surface which is certainly harder than simply reacting to a single price vector and reporting only the quantities demanded at that price. A failure to recognize the information processing constraints in the optimization problem is undoubtedly one of the reasons there has been limited success in using the theory of optimal auctions to explain the existence of pervasive institutions, such as the first- price sealed-bid auction used in competitive contracting or the posted price institution used in retailing.

Summary

Incentive compatibility captures the fundamental positivist notion of self-interested behaviour that underlies almost all economic theory and application. It has proven to be an organizing principle of great scope and power. Combined with the modern theory of mechanism design, it provides a framework in which to analyse such diverse topics as auctions, central planning, regulation of monopoly, transfer pricing, capital budgeting, and public enterprise management. Incentive compatibility provides a basic constraint on the possibilities for normative analysis. As such it serves as the

fundamental interface between what is desirable and what is possible in a theory of organizations.

See Also

- ▶ Efficient Allocation
- ▶ Externalities
- ▶ Lindahl Equilibrium
- ▶ Public Goods

Bibliography

Baron, D., and R. Myerson. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50: 911–930.

Dasgupta, P., P. Hammond, and E. Maskin. 1979. The implementation of social choice rules: Some general results on incentive compatibility. *Review of Economic Studies* 46: 185–216.

Gibbard, A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41: 587–602.

Groves, T., and J. Ledyard. 1977. Optimal allocation of public goods: A solution to the 'free rider' problem. *Econometrica* 45: 783–809.

Groves, T., and J. Ledyard. 1986. Incentive compatibility ten years later. In *Information, incentives, and economic mechanisms*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.

Harris, M., and A. Raviv. 1981. Allocation mechanisms and the design of auctions. *Econometrica* 49: 1477–1499.

Hurwicz, L. 1960. Optimality and informational efficiency in resource allocation processes. In *Mathematical methods in the social sciences*, ed. K. Arrow, S. Karlin, and P. Suppes, 27–46. Stanford: Stanford University Press.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and organization: A volume in honor of Jacob Marschak*, ed. R. Radner and C.B. McGuire, 297–336. Amsterdam: North-Holland.

Hurwicz, L. 1979. Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. *Review of Economic Studies* 46: 217–225.

Ledyard, J. 1978. Incomplete information and incentive compatibility. *Journal of Economic Theory* 18: 171–189.

Ledyard, J. and Roberts, J. 1974. *On the incentive problem with public goods*, Discussion paper, no. 116. Center for Mathematical Studies in Economics and Management Science, Northwestern University.

Lindahl, E. 1919. *Die Gerechtigkeit der Besteuerung*. Lund. Partial translation. In *Classics in the theory of public finance*, ed. R. A. Musgrave and A. T. Peacock. London: Macmillan, 1958.

Matthews, S. 1983. Selling to risk averse buyers with unobservable tastes. *Journal of Economic Theory* 30: 370–400.

Moore, J. and Repullo, R. 1986. *Subgame perfect implementation*, Working paper. London: London School of Economics.

Myerson, R.B. 1981. Optimal auction design. *Mathematics of Operations Research* 6: 58–73.

Myerson, R.B. 1985. Bayesian equilibrium and incentive compatibility: An introduction. In *Social goals and social organization: Essays in memory of Elisha Pazner*, ed. L. Hurwicz, D. Schmeidler, and H. Sonnenschein. Cambridge: Cambridge University Press.

Palfrey, T. and Srivastava, S. 1986. Implementation in exchange economies using refinements of Nash equilibrium. Graduate School of Industrial Administration, Carnegie-Mellon University, July 1986.

Palfrey, T., and S. Srivastava. 1987. On Bayesian implementable allocations. *Review of Economic Studies* 54: 193–208.

Postlewaite, A., and D. Schmeidler. 1986. Implementation in differential information economics. *Journal of Economic Theory* 39: 14–33.

Roberts, J., and A. Postlewaite. 1976. The incentives for price-taking behavior in large economies. *Econometrica* 44: 115–128.

Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Satterthwaite, M. 1975. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10: 187–217.

Schmeidler, D. 1980. Walrasian analysis via strategic outcome functions. *Econometrica* 48: 1585–1593.

Walker, M. 1981. A simple incentive compatible scheme for attaining Lindahl allocations. *Econometrica* 49: 65–71.

Incentive Contracts

Edward P. Lazear

Incentives are the essence of economics. The most basic concept, demand, considers how to induce a consumer to buy more of a particular good; that is, how to give him an incentive to purchase. Similarly, supply relationships are descriptions of how agents respond with more output or labour to additional compensation.

Incentive contracts arise because individuals love leisure. In order to induce them to forgo some leisure, or put alternatively, to put forth

effort, some form of compensation must be offered. The theme of this essay is that different forms of incentive contracts deal with some aspects of the problems better than others. The strength of one type of contract is the weakness of another. The labour market trades off these strengths and weaknesses and thereby selects a set of institutions. In what follows, the development of the literature on incentive contracts is briefly discussed. The emphasis is on concepts rather than specific papers or authors, so the bibliography is far from exhaustive.

To discuss incentive contracts, the most general concepts must be narrowed. This essay does that in two ways. First, attention here is restricted to the labour market. At a more general level, incentive contracts can relate to other areas as well. For example, the government may want to have a space satellite built at the lowest possible cost. To do so, incentives must be set appropriately or the producer may charge too much or fail to meet desired quality standards. This problem is analogous to those that arise in the labour context, but for the most part they are ignored, except when isomorphic with the labour market paradigm. Similarly, the law and economics literature is another area where incentive problems are studied, usually in the context of accident liability (see, for example, Green 1976; Polinsky 1980; Shavell 1980). These specific questions are ignored as well, except as they border on the labour market context. Second, the focus is on observability problems. Standard labour supply functions, where hours of work can be observed and paid, are incentive contracts. However, standard labour supply issues are eliminated from consideration since they are dealt with in other essays in *The New Palgrave*.

General Framework

An employer in a competitive environment must induce a worker to perform at the efficient level of effort or face extinction. The reason is simple: if one employer can, through clever use of an incentive contract, get a worker to perform at a more efficient level, that firm's cost will be lower. Lower costs imply that higher wages can be paid to workers and all workers will be stolen from inefficient firms. As a result, the objective function that is taken as standard for the firm is:

$$\underset{F}{\text{Max}}\ F(Q,E) - C(E), \tag{1}$$

where Q is output and E is worker effort. Thus $F(Q, E)$ is the compensation schedule that the firm announces to the worker; $C(E)$ is the worker's cost of effort function, to be thought of as the dollar cost associated with supplying effort level E.

The competitive nature of the firm in factor and product markets implies that the firm must maximize worker net wealth as in (1) subject to the zero profit constraint:

$$Q = F(Q,E). \tag{2}$$

Output is defined so that each unit sells for \$1 (the numeraire). Thus (2) merely says that output, Q, must be paid entirely to the worker otherwise another firm could steal the worker away by paying more.

The incentive problem arises because the worker takes the compensation scheme $F(Q, E)$ as given and chooses effort to maximize expected utility. Once the worker has accepted the job, his problem is:

$$\underset{F}{\text{Max}}\ F(Q,E) - C(E). \tag{3}$$

The worker's effort supply function comes from solving the first-order condition associated with (3) or

$$C'(E) = \frac{\partial F}{\partial Q}\cdot\frac{\partial Q}{\partial E} + \frac{\partial F}{\partial E}, \tag{4}$$

which says that the worker sets the marginal cost of effort equal to its marginal return to him. The transformation of effort into output, (i.e. $\partial Q/\partial E$) depends on the production function. A convenient specification is

$$Q = E + v, \tag{5}$$

so that output is the sum of effort, E, and luck, v.

An incentive contract selects $F(Q, E)$ subject to the zero-profit constraint, (2), taking into account that the worker behaves according to (4). There are an infinite variety of incentive contracts that are subsumed by $F(Q, E)$. To make things clear, we consider two polar extremes – the salary and the piece rate (for a more detailed treatment, see Lazear 1986).

Let us define a salary as compensation that depends only on input so that $F(Q, E)$ takes the form $S(E)$. An hourly wage is an example. Irrespective of the amount that is produced during the hour, the worker receives a fixed amount that depends only on the fact that he supplies E of effort for the hour. (Of course, difficulty in measuring E may be a compelling reason to avoid this form of incentive contract.) At the other extreme is a piece rate where compensation depends only on output so that $F(Q, E)$ takes the form of $R(Q)$. There, no matter how much or how little effort the worker exerts, his compensation depends only on the number of units produced. Both salaries and piece rates are incentive contracts; the first provides incentives by paying workers on the basis of input. The second provides incentives by paying on the basis of output. More sophisticated incentive contracts, which blend the two or use multiperiod approaches are discussed later.

The Principal–Agent Problem

At the centre of the incentive contract literature is the 'principal–agent' problem. The principal, say, an employer, wants to induce its agent, say, a worker, to behave in a way that is beneficial to the employer. The problem is that the principal's knowledge is imperfect; either he cannot see what the agent does (as in the case of a taxi driver who can sleep on the job) or he cannot interpret the actions (as in the case of an auto mechanic who replaces a number of parts to correct a perhaps simple malfunction). The incentive contracts that can be used to address the problem were discussed early by Ross (1973), Mirrlees (1976), Calvo and Wellisz (1978) and by Becker and Stigler (1974). The last in particular, uses a sampling approach. For example, a politician can be required to post a large bond on taking office. If he is caught engaging in some malfeasant behaviour, he forfeits the bond. This contract is based on output, which is observed infrequently or imperfectly. Other kinds of incentive contracts are discussed in the following sections.

Payment by Output

Sharecropping

One of the earliest examples of incentive contracts that is based on output is sharecropping. In sharecropping, the owner contracts to split the output of the land in some proportion with the individual who farms and lives on it. It was also one of the first incentive schemes that was clearly analysed (see Johnson 1950, and later Cheung 1969, and Stiglitz 1974). The original problem as formulated in sharecropping can be seen as follows.

Payment is conditional only on Q and by some fixed proportion so that the worker receives γQ. Using (4) and (5), compensation of this sort implies that the worker's first-order condition is

$$C'(E) = \gamma$$

so that the worker sets the marginal cost of effort equal to γ. But (5) implies that the marginal value of effort is \$1, which exceeds γ so that the worker puts forth too little effort. This is inefficient. Additionally, if the farmer can obtain land without limit, he pushes his sharecropping acreage to the point where the next unit of land has zero marginal product. This is clearly inefficient but can be remedied if landowners can select sharecroppers and terms according to the amount of land each works. Both the owner and worker could be made better off if the worker could be induced, by another incentive contract, to produce where $C'(E) = 1$.

Renting the land to the farmer and allowing the farmer to keep all of the output accomplishes this. Under rental, the worker's compensation is [Q-Rent]. By (4) and (5), the worker is induced to set $C'(E) = 1$; the marginal cost and marginal

value of output are equated. Of course, rental does not solve all of the problems. Absent in the production function in (5) is that maintenance may be required. For example, if the farmer does not fertilize the land, it may not produce as well in the future. A renter, who can move on to the next plot after the soil is drained of minerals, has little incentive to put resources into the land. Thus the solution is to sell the land to the farmer. Then the individual who works the land has the correct incentives, either because he will continue to use it in the future or because the sale price will reflect the quality of the land. But sale of the land begs most of the questions. The sale may not come about because of the farmer's capital constraints, because of his lack of entrepreneurial skill, or because of his distaste for risk. (Note that risk is shifted from owners to farmers even in sharecropping and renting. Only labour contracts based exclusively on effort shift the risk entirely to the owner.)

The sharecropping paradigm applies to industrial production as well. Profit-sharing arrangements are, in many respects, like sharecropping. This is especially true when there is only one worker. Partnerships are similar. The same incentive problems arise. A worker who can quit and move on to another firm without penalty does not have the same desire to maintain the equipment as the firm's owner. Again the solution is to sell the capital to the worker, but this simply redefines the owner. Then there is no principal-agent problem because there is no agent. This can be considered in more detail in the next section.

Piece Rates

Piece-rate compensation is not much different from sharecropping, the latter being a special case of the former (see Stiglitz 1975). The owner allows the worker (or farmer) to use his capital (or land) and pays the worker according to some function of output. In the simplest scheme, a linear piece rate is used and the worker is paid rate R per unit Q so that compensation is RQ. The worker's maximization problem (3) and (4) implies that the worker sets $C'(E) = R$. The firm's zero-profit constraint in (2) implies that $Q = RQ$ or that $R = 1$. Thus the piece rate is efficient because the worker sets the marginal cost of effort equal to its marginal social value, \$1.

The issue is only slightly more complicated if capital is involved. A linear piece rate with an intercept (i.e. compensation equal to $A + RQ$) will do the job. This incentive contract achieves first-best efficiency. The worker's first-order condition, (4), still guarantees that he sets $C'(E) = R$. The intercept drops out. But the zero-profit constraint now becomes:

$$Q - \text{rental cost of capital} = A + RQ.$$

The firm must 'charge' the worker for the cost of using the capital, but how should this be done? R can be reduced below 1 or A can be set to a negative number. The answer is that $A = -$(rental cost of capital) and $R = 1$. Since (4) does not contain A, the worker does not respond to changes in A. However, reducing R below 1 causes the worker to reduce effort. Thus the efficient incentive contract, which also maximizes worker wealth subject to the firm's zero-profit constraint, requires that $R = 1$. Zero profit requires that $A = -$ (rental cost of capital).

A major advantage to the use of piece rates as an incentive contract is that it tolerates heterogeneity of worker ability. More able – that is, lower effort cost – workers choose higher levels of effort but are paid more. There is no inefficiency involved in having workers of both types in the firm. Of course, if capital is important so that the worker is 'charged' A for the right to work on a machine, only workers above some threshold ability level will choose to work. But workers self-sort. There is no need for the firm to do anything other than pay the efficient piece rate, in this case $R = 1$.

Linear piece rates are no longer appropriate incentive contracts if workers are risk-averse. In general, a non-linear scheme will do better but will fail to achieve first-best solutions. As long as asymmetric information exists, so that individual actions cannot be observed and contracted upon, Pareto optimal risk-sharing is precluded (see Hölmstrom 1979; Harris and Raviv 1979).

Payment of Relative Output

The study of relative compensation has become increasingly important. There are two approaches in this literature. The first, from Lazear and Rosen (1981), characterizes the labour market as a tournament, where one worker is pitted against another. The one with the highest level of output receives the winning prize (i.e. the high-wage job) while the other gets the losing prize (i.e. the low-wage job). By increasing the spread between the winning and losing prizes, incentives are provided to work hard. The optimum spread induces workers to move to the point where the marginal cost of effort exactly equals the marginal (social) return to it. The major advantages to payment by tournament method are twofold. First, tournaments require only that relative comparisons be made. It may be cheaper to observe that one worker produces more than another than to determine the actual amount that each produces. Second, compensation by rank 'differences out' common noise. For example, sales may be low because the economy is in a slump, which has nothing to do with worker effort. Risk aversion operates against penalizing or rewarding workers for factors over which they have no control. But since the slump affects both workers equally, relative comparisons are unaffected. The best worker still produces more, even though both produce small amounts.

Tournament-type incentive contracts induce workers to behave efficiently if they are risk neutral. They are easy to use but carry one major disadvantage. Workers increase the probability of winning, not only by doing well themselves but also by causing the opponent to do poorly. Thus tournaments discourage cooperation. This results in wage compression, which works to discourage the aggressive behaviour of workers who are competing for the same job. Other work in the area of tournament-type incentive contracts includes Nalebuff and Stiglitz (1983), Green and Stokey (1983) and Carmichael (1983).

The second approach, from Hölmstrom (1982), suggests that if levels of output can be observed, then payments can be based, at least in part, on a team average. As Hölmstrom points out, a tournament is not a sufficient statistic, so that using a team average allows the firm to better address risk aversion. This incentive device also takes out common noise. A peer average picks up disturbances that are common to the industry and allows the firm to cater to the tastes of risk-averse workers.

Payment by Input

Observability of Effort

It is commonly alleged that payment of a salary or hourly wage does not provide workers with the appropriate incentives. Whether or not this is true depends on the connection between the measurement of time and measurement of effort. To see this, suppose that effort can be observed perfectly, but that output cannot be observed at all. For example, suppose that it is easy to measure the number of calories burned up by a worker during his work day, but it is impossible to separate his output from that of his peers. Payment by effort is a first-best incentive contract. The compensation scheme that pays the worker \$1 per unit of effort exerted induces him to set $C'(E) = 1$, which, as we have seen, is first best. Note further that this is first best even for risk-averse workers since compensation does not vary with random productivity shocks, v (see Hall and Lilien 1979).

The allegation that effort pay does not provide incentives is based on the difference between hours of work and effort. If hours were a perfect proxy for effort, then payment of an hourly wage would be an optimal incentive contract. But because workers can vary work per hour, the connection breaks down. Payment per hour provides appropriate incentives for choice of the number of hours, but does not deal with what is done within the hour.

Payment by Effort and Worker Sorting

Piece rates induce workers to sort appropriately. Above, it was argued that workers who cannot produce a sufficiently high level of output will not come to a firm that 'charges' for use of capital. Salaries (or hourly wages) that pay on the basis of

an imperfect measure of effort encourage the lower-quality workers to come to the firm. Lazear (1986) demonstrates that a separating equilibrium (see, e.g., Rothschild and Stiglitz 1976; Salop and Salop 1976) exists where high-quality workers choose to work at firms that pay piece rates and low-quality ones choose salaries. The difference in quality across firms might lead one to conclude that movement to output-based incentive contracts increases total output. In fact, the reverse may well be true. In the same sense that screening in Spence (1973) is socially unproductive, forcing salary firms to adopt piece-rate incentive contracts wastes resources on a potentially useless signal.

Incentive Contracts and Product Quality

Sometimes quantity is easier to observe than quality. The problem with incentive contracts that are based on output quantity is that they induce the worker to go for speed and to ignore quality. If quality can be observed, then the worker can be compensated appropriately for quantity and quality. The appropriate compensation function is essentially the consumer's demand for the product as it varies with quality and quantity. But if quality cannot be observed, payment by input 'solves' the quantity/quality problem. If the worker is paid, say, by hour, and is merely instructed to produce goods of a given quality, he has no incentive to deviate from that instruction. Compensation is based only on input, so there is no desire to rush the job. Of course, this requires a method of monitoring effort cheaply (see Lazear 1986, for a full discussion of the trade-offs).

Other Issues in Incentive Contracting

Efficient Separation and Long-Term Investments

A properly structured incentive contract must induce the correct amount of long-term investment. The problem is most clearly seen in the context of specific human capital, as in Becker (1962, 1975). Specific human capital is only valuable when the worker is employed at the current firm. As such, workers are reluctant to invest in specific capital because the firm may capriciously fire the worker, in which case the investment is lost. Similarly, firms are reluctant to invest because the worker may capriciously quit. The incentive contract that Becker suggests is a sharing of investment costs and returns by both workers and firms (Hashimoto and Yu 1980, model this more precisely). Kennan (1979) points out that a particular kind of severance pay solves the investment problem. It is akin to the liability rules that are efficient in auto accident problems. But as Hall and Lazear (1984) argue, these rules may actually induce too much investment. Since a worker is compensated for the full investment whether work occurs or not, he has no incentive to account for situations that make a separation optimal. For example, if it were optimal to sever the work relationship 25 per cent of the time, the worker should behave as if a specific investment that yields $1 return only yields $0.75. A full-reimbursement severance pay arrangement ensures a full $1, irrespective of the status of work, and induces too much investment.

More general issues of efficient separation arise in the labour market context, and incentive contracts must be structured to deal with these problems. Hall and Lazear (1984) consider a variety of different incentive contracts and conclude that none generally achieves first best. One that comes close to doing so is Vickrey's (1961) bilateral auction approach. There, compensation and work are separated so that the worker and firm have incentives to reveal the true relevant values. Another scheme is coordinated severance pay, suggested by d'Aspremont and Gerard-Varet (1979). Sufficiently high penalties on the firm associated with a worker's refusal to work induces the firm to behave in a manner that is apparently first best.

Intertemporal Incentive Contracts

Sometimes, the fact that workers live for more than one period allows contracts to be structured in a way that solves incentive problems. This is the subject of Lazear (1979, 1981). The problem is that as a worker approaches the end of his career, he has an incentive to shirk because the costs, even of being fired, are reduced as his retirement date draws near. A way to discourage shirking is to

tilt the age-earnings profile and couple it with a contingent pension. Young workers are paid less than their marginal products; old workers are paid more. In equilibrium, shirking is discouraged and workers receive exactly their lifetime marginal products. The distortion in the timing of the payments implies that workers do not voluntarily choose to work the correct number of hours. Thus hours constraints are required, an extreme form of which is mandatory retirement. Other work that has refined or provided empirical support for that concept is Kuhn (1986) and Hutchens (1986a, b).

There are other papers that focus on the intertemporal aspects of incentive contracts. The first, Fama (1980) argues that the market provides a discipline on workers. In a spot market, the wage that another firm is willing to offer a worker next period depends on how well he did last period. Fama shows that this can act as a perfect incentive device. Of course, no end-game problems are addressed by this mechanism, but it does demonstrate the possibility of incentive provision even without explicit or implicit contracts. The second idea is attributable to Rogerson (1985). The emphasis here is on risk-sharing, but the work has some features in common with Fama (1980). In particular, memory plays a strong role in these incentive contracts, so that an outcome that affects the current wage also affects the future wage.

Intertemporal Strategic Behaviour by Firms

Once intertemporal contracts are considered, it is necessary to examine the issue of opportunistic behaviour by firms. It may be that a firm does not know a worker's cost of effort function, $C(E)$. Actions that the worker takes may reveal information about that function. The firm can use that information in subsequent periods against the worker. As a result, the worker attempts to disguise $C(E)$, leading to inefficiencies. Such is the case of salesmen, whose next period quota depends on this period's performance. In Lazear (1986) it is shown that a properly structured contract in a competitive labour market can undo the effects of this kind of strategic behaviour. This is a specific example of the general theorem on revelation presented in Harris and Townsend (1981). It is also related to the literature on planned economies, since bureaucrats tend to make things look worse than they are to lessen next period's requirements or to increase next period's budget allocation (see, e.g., Weitzman 1976, 1980; Fan 1975).

Insurance

Finally, there is a closely related literature that examines insurance contracts. That literature focuses, for the most part, on the trade-off between insurance and efficiency in the labour market. Some of the more important papers in that literature include Harris and Hölmstrom (1982), Grossman and Hart (1983) and Green and Kahn (1983).

Conclusion

Although incentive problems are pervasive, the market has found a number of solutions. These involve payment by output of the piece rate or sharecropping variety; payment by relative output, exemplified by labour market tournaments; payment by measured input, such as hours of work; and multi-period incentive contracts. The contracts do not always achieve the first best, especially when risk aversion is an issue. Still, the rich variety of institutions that address incentive problems and the large amount of literature devoted to study attest to the problem's importance in the labour market context.

See Also

- ▶ Implicit Contracts
- ▶ Incomplete Contracts
- ▶ Layoffs
- ▶ Principal and Agent (i)

Bibliography

Becker, G.S. 1962. Investment in human capital: A theoretical analysis. *Journal of Political Economy* 70: 9–49.

Becker, G.S. 1975. *Human capital: A theoretical and empirical analysis, with special reference to education*, 2nd ed. New York: Columbia University Press for the National Bureau of Economic Research.

Becker, G.S., and G.J. Stigler. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3: 1–18.

Calvo, G., and S. Wellisz. 1978. Supervision, loss of control and optimum size of the firm. *Journal of Political Economy* 86: 943–952.

Carmichael, H.L. 1983. The agent–agents problem: Payment by relative output. *Journal of Labor Economics* 1: 50–65.

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.

d'Aspremont, C., and L.A. Gerard-Varet. 1979. Incentives and incomplete information. *Journal of Public Economics* 11: 25–45.

Fama, E. 1980. Agency problems and the theory of the firm. *Journal of Political Economy* 88: 288–307.

Fan, L.-S. 1975. On the reward system. *American Economic Review* 65: 226–229.

Green, J.R. 1976. On the optimal structure of liability laws. *Bell Journal of Economics* 7: 553–574.

Green, J.R., and C. Kahn. 1983. Wage employment contracts. *Quarterly Journal of Economics* 98: 173–188.

Green, J.R., and N.L. Stokey. 1983. A comparison of tournaments and contracts. *Journal of Political Economy* 91: 349–364.

Grossman, S., and O. Hart. 1983. Implicit contracts under asymmetric information. *Quarterly Journal of Economics* 71: 123–157.

Hall, R.E., and E.P. Lazear. 1984. The excess sensitivity of layoffs and quits to demand. *Journal of Labor Economics* 2: 233–258.

Hall, R.E., and D. Lilien. 1979. Efficient wage bargains under uncertain supply and demand. *American Economic Review* 69: 868–879.

Harris, M., and B. Hölmstrom. 1982. A theory of wage dynamics. *Review of Economic Studies* 49: 315–333.

Harris, M., and A. Raviv. 1979. Optimal incentive contracts with imperfect information. *Journal of Economic Theory* 20(2): 231–259.

Harris, M., and R. Townsend. 1981. Resource allocation under asymmetric information. *Econometrica* 49: 33–64.

Hashimoto, M., and B. Yu. 1980. Specific capital, employment contracts, and wage rigidity. *Bell Journal of Economics* 536–549.

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

Hölmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13: 324–340.

Hutchens, R. 1986a. Delayed payment contracts and a firm's propensity to hire older workers. *Journal of Labor Economics* 4(4): 439–457.

Hutchens, R. 1986b. An empirical test of Lazear's theory of delayed payment contracts. Working paper, Cornell University Institute for Labor and Industrial Relations.

Johnson, D.G. 1950. Resource allocation under share contracts. *Journal of Political Economy* 58: 111–123.

Kennan, J. 1979. Bonding and the enforcement of labor contracts. *Economic Letters* 3: 61–66.

Kuhn, P.J. 1986. Wages, effort, and incentive compatibility in life-cycle employment contracts. *Journal of Labor Economics* 4: 28–49.

Lazear, E.P. 1979. Why is there mandatory retirement? *Journal of Political Economy* 87: 1261–1284.

Lazear, E.P. 1981. Agency, earnings profiles, productivity and hours restrictions. *American Economic Review* 71: 606–620.

Lazear, E.P. 1986. Salaries and piece rates. *Journal of Business* 59(3): 405–431.

Lazear, E.P., and S. Rosen. 1981. Rank order tournaments as optimum labor contracts. *Journal of Political Economy* 89: 841–864.

Mirrlees, J.A. 1976. The optimal structure of incentives with authority within an organization. *Bell Journal of Economics* 7: 105–131.

Nalebuff, B.J., and J.E. Stiglitz. 1983. Prizes and incentives: Toward a general theory of compensation and competition. *Bell Journal of Economics* 14: 21–43.

Polinsky, A.M. 1980. Strict liability vs. negligence in a market setting. *American Economic Review* 70(2): 363–367.

Rogerson, W.P. 1985. Repeated moral hazard. *Econometrica* 53: 69–76.

Ross, S.A. 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, J., and S. Salop. 1976. Self-selection and turnover in the labor market. *Quarterly Journal of Economics* 90: 619–627.

Shavell, S. 1980. Strict liability versus negligence. *Journal of Legal Studies* 1–25.

Spence, A.M. 1973. Job market signalling. *Quarterly Journal of Economics* 87: 355–374.

Stiglitz, J.E. 1974. Incentive 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 and Management Science* 6: 552–579.

Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16: 8–37.

Weitzman, M. 1976. The new Soviet incentive model. *Bell Journal of Economics* 7(1): 251–257.

Weitzman, M. 1980. The 'Ratchet Principle' and performance incentives. *Bell Journal of Economics* 11(1): 302–308.

Income

D. Usher

Like 'supply', 'demand', 'rent', 'welfare' and 'utility', the word 'income' is a part of common speech that has entered economics as a technical term. *The Concise Oxford Dictionary* defines income as 'receipts from one's lands, work, investment etc'. That meaning carries over into economic theory, where, for instance, a consumer may be said to maximize utility subject to an income constraint or a firm may be said to maximize income accruing to its stockholders. The meaning of income is somewhat modified in the construction of income statistics. These are employed in two quite distinct contexts: as the basis for income taxation and as generalized to national income.

In each context, the definition of income is governed by the purpose of the statistics. Personal and corporate income are defined to serve as criteria for taxing people and corporations, the main principles behind the definitions being equity or fairness among people with different sources of income and efficiency in the economy as a whole. The purpose of income within the national accounts is less easily defined. The national accounts are an intricate set of statistics intended to describe the economy as a whole, primarily but not exclusively to facilitate counter-cyclical policy. The simple concept of income as applied to a person is extended to the entire nation in several ways: national income is the sum of the earnings of all factors of production; national product is the value of output of all goods and services; national expenditure is the sum of each person's expenditure on goods and services. All three would be equal in a world without depreciation, indirect taxation or subsidies. Income statistics also serve as the basis for comparisons among regions, products and occupations; converted to real income, they become the basis for the measurement of the rate of economic growth.

As a first approximation, we may say that national income is the sum of all personal incomes, but the definition of income for tax purposes differs in several important respects from the definition in the national accounts. The major differences can be classified under the headings of scope, intermediation and timing.

The scope of income is a trade-off between two objectives, to include all benefits to consumers as part of income, even benefits arising from non-market activity, and to construct statistics that are reasonably precise and beyond dispute. The latter consideration is relatively more important for tax purposes. Thus money values of the services of owner-occupied housing, food grown and consumed on farms, and direct provision of food and lodging for the armed services, are usually included in national income but almost never as part of the tax base. Housework, on the other hand, is included in neither definition, giving rise to the old paradox that the national income falls if a man marries his housekeeper. There is an ongoing debate in public finance as to whether the base for personal taxation should be income as a whole (consumption plus investment) or just consumption.

The major issue in the timing of income concerns capital gains which are sometimes included in income for tax purposes but are never part of the national income. They are excluded from the national income so that beneficial changes in technology and other aspects of the economy appear as part of income in the years when they materialize as goods and services rather than when they are first anticipated. Personal income is another matter. A person becomes wealthy in the year his assets appreciate, regardless of the dates of the increases in the marginal products of the corresponding capital goods. The usual justification for including capital gains as part of taxable income is that a person should be taxed when be becomes wealthy, just as he is taxed when he earns ordinary income.

The inclusion of capital gains in the tax base creates several problems: assets may appreciate as part of a general inflation; taxation of capital gains may only be feasible when gains are realized

rather than when they accrue; gain on human capital is automatically exempt from the tax; taxation of capital gain may be double taxation of the return to capital because future tax on the earnings to capital goods is discounted back in the price at which the goods are sold. However, the exclusion of capital gains from the tax base creates an incentive for firms to seek ways to disburse money to people as non-taxable capital gains rather that as taxable income.

Both personal and national income are defined net of the cost of intermediate products. The income of the travelling salesman is net of the cost of his car, and the national income is net of the aggregate cost of transport for business purposes – of the cost of haulage of goods, business travel, and so on. The boundary between final and intermediate products is sometimes problematic. For example, it is not always clear how to classify the business lunch, especially for participants who would rather diet. Current expenditure of government is classified in the national accounts as consumption, yet it is arguable that most such expenditure is either intermediate product (social overhead for the economy as a whole) or intangible investment (the obvious instance being current expenditure on research).

Depreciation is like an intermediate product. An input to production that is used up in the course of the year or incorporated into output is unambiguously intermediate, and its cost is excluded on that account from the measure of income. An input to production that lasts more than a year but depreciates somewhat during the first year is financially equivalent to the sum of an ordinary intermediate product and an input that only becomes available at the end of the first year. To deduct depreciation from income is to treat the intermediate component of investment like an ordinary intermediate good. On the other hand, it is often difficult in practice to determine what depreciation ought to be. The tax code specifies rates of allowable depreciation for each type of capital equipment, for it is more important in this context to be precise and predictable than to be right. There is more flexibility in the national accounts and an attempt is made to measure the true loss over the year in the value of capital goods. Loss of value may be deterioration or obsolescence. Both belong as part of depreciation, especially when obsolescence is anticipated, for it makes no difference at the time a machine is purchased whether it is destined to deteriorate through use or to become worthless as better machines are developed. Unanticipated obsolescence is more problematic, for there is something anomalous in reducing this year's national income for the fall in the value of capital goods brought about by expected technical change next year when the benefit of the change itself is excluded. Income would fall though there would be no reduction of output in the current year and people would be better off in the long run.

See Also

- ▶ Capital Gains and Losses
- ▶ Depreciation
- ▶ National Income
- ▶ Real Income
- ▶ Social Accounting

Bibliography

On the definition of personal income, see Simons 1938. The classic discussion of the concept of national income, related conceptual problems and the history of the development of the national accounts is Studenski 1961. Most countries produce national accounts and publish details of how they are compiled. For a proposed standardization, see *A System of National Accounts*, United Nations, 1968. Current research in national accounting is likely to appear in *The Review of Income and Wealth*.

Simons, H. 1938. *Personal income taxation: The definition of income as a problem of fiscal policy*. Chicago: University of Chicago Press.

Studenski, P. 1961. *The income of nations*. New York: New York University Press.

United Nations. 1968. *A system of national accounts*. New York: United Nations, Department of Economic and Social Affairs, Studies in methods, series F, no. 2.

Income Mobility

Gary S. Fields

Abstract

Income mobility means different things to different people. This article explains the six different mobility concepts used in the literature, reviews the various indices used in the mobility literature to measure these concepts, summarizes the difference the use of different mobility concepts and measures makes in practice, presents the axiomatic approach to income mobility, and discusses a number of other issues that arise in the mobility literature.

Keywords

Absolute mobility; Exchange mobility; Income mobility; Inequality (measurement); Intergenerational income mobility; Movement studies; Relative mobility; Structural mobility; Time-independence studies; Translation invariance

JEL Classifications

J62; C43; D31; J3; J6

What is income mobility? Extensive surveys of the income and earnings mobility literatures may be found in Atkinson et al. (1992), Maasoumi (1998), Solon (1999), and Fields and Ok (1999a). ('Income' refers to income from all sources while 'earnings' refers to income earned in the labour market.) Mobility analysts agree on one defining feature: 'income mobility' is about how much income each recipient receives at two or more points in time. In this way, income mobility studies are distinguished from studies of the inequality and poverty aspects of income distribution, both of which are based (typically) on anonymous cross sections or (less frequently) marginal distributions of the joint distributions.

The following notation is used throughout this article. Let $x = (x^1, \ldots, x^n)$ denote a vector of 'incomes' in an initial year. This vector is 'personalized' in the sense that the same recipient units are followed over time. It is conventional to array the recipients in the base year from lowest income to highest. Whether this convention is followed or not, it is essential to keep the same order for subsequent years (or generations). Denote the ordered vector in a subsequent year by $y = (y^1, \ldots, y^n)$. The micro-mobility data, also termed in the literature the pattern of 'distributional change', is summarized by the transformation $x \rightarrow y$ in the two-period case or more generally the transformation $x \rightarrow y \rightarrow z \rightarrow \cdots$ in the T-period case. The extent of mobility associated with the transformation $x \rightarrow y$ will be denoted by $m(x, y)$.

Beyond agreeing that income mobility studies are about transformations of the type $x \rightarrow y$ or $x \rightarrow y \rightarrow z \rightarrow \cdots$, the literature is marked by considerable disagreement. This is because the term 'income mobility' connotes precise but *different* ideas to different researchers. It is for this reason that mobility analysts often have trouble communicating with each other, with other social scientists, or with the general public. Furthermore, these differences in notions of what income mobility is remain even after agreement is reached on a number of other aspects of the mobility under consideration. These other aspects, discussed in the following paragraphs, are whether the context is intergenerational or intragenerational, what the indicator of social or economic status is, and whether the analysis is at the macro-mobility or micro-mobility level.

One issue is whether the aspect of mobility of interest is intergenerational or intragenerational. In the *intergenerational* context, the recipient unit is the family, specifically a parent and a child. In the *intragenerational* context, the recipient unit is the individual or family at two different dates. The issues discussed in this article apply equally to both.

Second, agreement must be reached on an indicator of social or economic status and the

choice of recipient unit. For brevity, I shall talk about mobility of 'income' among 'individuals'.

Third, the mobility questions asked and our knowledge about mobility phenomena may be grouped into two categories, macro and micro. *Macro*-mobility studies start with the question, 'How much economic mobility is there?' Answers are of the type '*a* per cent of the people stay in the same income quintile', '*b* per cent of the people moved up at least $1,000 while *c* per cent of the people moved down at least $1,000', 'the mean absolute value of income change was $*d*,' and 'in a panel of length T, the mean number of years in poverty is t^*.' The macro-mobility studies often go beyond this question to ask, 'Is economic mobility higher here than there and what accounts for the difference?' Answers would be of the type, 'economic mobility has been rising over time', '*A* has more upward mobility than *B* because economic growth was higher in *A* than in *B*', and 'incomes are more stable in *C* than in *D* because *C* has a better social safety net'. *Micro*-mobility studies, on the other hand, start with the question, 'What are the correlates and determinants of the income or positional changes of individual income recipients?' The answers to these questions would be of the type, 'unconditionally, income changes are higher for the better-educated' and 'other things equal, higher initial income is associated with lower subsequent income growth'.

These three issues – intergenerational versus intragenerational, changes in the distribution of what among whom, and macro-mobility versus micro-mobility – help determine which kind of mobility analysis is being undertaken. Yet major differences remain. It is to these that we now turn.

Mobility Concepts and Measures

At least 20 mobility measures have been used in the literature. Many empirical mobility studies divide base- and final-year incomes into quantiles (for example, quintiles or deciles) and calculate immobility ratios, mean upward movements, and the like (Fields 2001). Other studies estimate correlation coefficients between base-year and final-year incomes (Atkinson et al. 1992). In the intergenerational mobility literature, it is common to calculate intergenerational elasticities, that is, the coefficient obtained when the logarithm of the child's income is regressed on the logarithm of the parent's (Solon 1999).

In each case, we may ask, what are the various measures measuring? The essential answer is this: *different indices measure different underlying entities*. Whenever one of these underlying entities is measured, other information contained in the joint distribution of initial and final incomes is lost.

What are the different underlying entities that the various income mobility measures measure? The first distinction to be drawn is between measures of time independence and measures of movement. The question asked by *time-independence* studies is, how dependent is current income on past income? One commonly used measure of time independence is the beta coefficient commonly calculated in the intergenerational mobility literature by regressing the log-income of the child on the log-income of the parent.

Movement studies ask a different question, namely: in comparisons of incomes of the same individuals between one year and another, or of parents and children between one generation and another, how much income movement has taken place? The various movement indices in the literature may usefully be classified into five categories or concepts ('concepts' because they are different underlying entities, not alternative measures of the same underlying entity).

Positional movement (or 'quantile movement') is about the movement of individuals among various positions (quintiles, deciles, centiles, or ranks) in the income distribution. An individual experiences positional movement if and only if he or she changes quintiles, deciles, centiles, or ranks. Positional movement in a population is greater the more such positional changes there are and/or the larger these positional changes are. King (1983) derived a broad class of positional movement indices axiomatically, one member of which is

$$M_K(x,y) = 1 - \exp\left[-\frac{\gamma}{n}\sum_{i=1}^{n}\frac{|z_i - y_i|}{\mu(y)}\right],$$

where γ is the observer's degree of immobility aversion, z_i is the income level agent i would have obtained if his or her rank order did not change during the process $x \rightarrow y$, and $\mu(y)$ is the mean income in distribution y.

Like positional movement, *share movement* is relative but it is relative in a different way. Share movement takes place if and only if an individual's income rises or falls relative to the mean. Thus, an individual can experience upward or downward share movement even if his or her income in dollars is unchanged and/or if he or she does not change position within the income distribution. Share movement in the population reflects the frequency and magnitude of these individual share changes. One attractive index of share movement in a population is the mean absolute value of share changes

$$M_S(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i}{\mu_y} - \frac{x_i}{\mu_x}\right|,$$

where $\mu(x)$ and $\mu(y)$ are the means of distributions x and y respectively.

Another concept is *non-directional income movement* (also called '*flux*'), which gauges the extent of fluctuation in individuals' incomes. To illustrate, suppose that in a two-person economy one person's income goes up by $10,000 while another's goes down by $10,000. Those who see an average income change of $10,000 are non-directional income movement adherents. Two indices of non-directional income movement have been suggested by Fields and Ok (1996, 1999b):

$$M_{F-O_1}(x,y) = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

and

$$M_{F-O_2}(x,y) = \frac{1}{n}\sum_{i=1}^{n}|\log y_i - \log x_i|.$$

Suppose, however, that, when one person's income goes up by $10,000 and another's goes down by $10,000, the observer cares not only about the amounts of the income changes but also about their direction. *Directional income movement* may be judged using a linear or a concave valuation function. One valuation function which embodies concavity is the mean change in log-incomes (Fields and Ok 1999a, b):

$$M_{F-O_3}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log x_i).$$

As a fifth and final notion of income movement, consider how the income changes experienced by individuals cause the inequality of longer-term incomes to differ from the inequality of base-year incomes. *Mobility as an equalizer of longer-term incomes* would judge that a pattern of income change (1, 3) → (1, 5) would *disequalize* longer-term income relative to the base, while a pattern of income change (1,3)→ (5,1) would *equalize* longer-term income relative to the base. This concept is well-established in the literature (Schumpeter 1955; Shorrocks 1978b; Atkinson et al. 1992; Slemrod 1992; Krugman 1992; Jarvis and Jenkins 1998), but only recently has a class of measures of this concept been proposed (Fields 2005). One family within this class is

$$\varepsilon \equiv 1 - (I(a)/I(x)),$$

where x is the vector of base-year incomes, y is the vector of final-year incomes, a is the vector of average incomes, the i'th element of which is $a^i \equiv \frac{x^i + y^i}{2}$, and $I(.)$ is a cross-sectional inequality measure such as the Gini coefficient or the Theil index.

We thus have six mobility concepts and a large number of measures. Because these concepts are fundamentally different from one another, it is important for analysts to choose the concepts that are of greatest interest to them and then measure those concepts. Let us now turn to a brief empirical review of studies that have used two or more of these concepts.

Different Mobility Concepts in Practice

The previous section distinguished between time independence, positional movement, share movement, non-directional income movement, directional income movement, and mobility as an equalizer of longer-term incomes. How do these six concepts and the measures of them compare in empirical work? Specifically, which country has more mobility than another? Has mobility been rising or falling over time within a country? Are some groups in the population more or less mobile than others?

The answers to these questions have been shown empirically to depend on which mobility concept is used. In comparing *OECD countries*, some countries were found to be more mobile than others with the use of measures of some concepts and less mobile than others with the use of measures of other concepts (OECD 1996; 1997). When we looked over time, in the *United States* measures of four concepts (time independence, positional movement, share movement, and income flux) all peaked in 1980–5 but measures of two other concepts did not: directional income movement exhibits a saw-tooth pattern, while mobility as an equalizer of longer-term incomes exhibits a peak followed by a valley (Fields et al. 2002; Fields 2005). In *France*, mobility differences among demographic groups have been explored (Buchinsky et al. 2004). The answers to the questions 'Who has more mobility: women or men? Better-educated or less-educated workers?' were shown to differ depending on which mobility concept was used. By gender, women in France have *more* time independence and positional movement than men, *less* share movement than men, *about the same* non-directional and directional movement in logs, and *about the same* amount of mobility as an equalizer of longer-term incomes. By education, those with the highest educational attainments have *less* time independence and positional movement, and if anything *more* share movement, flux, and directional income movement in logs. In *Argentina*, too, measures of the six different concepts produced qualitatively different results (Sánchez Puerta 2005).

Looking at changes over time, some mobility indices increased, some decreased, and some showed no clear trend. Comparing population subgroups (genders, educational levels, age ranges, regions, initial quintiles, and initial sector), some groups were found to have higher earnings mobility for some concepts and lower earnings mobility for others; no group was found to have higher mobility than others for every mobility concept. Finally, in both *Venezuela* and *Mexico*, the time trend of mobility was found to vary according to the notion of mobility measured (Freije 2001; Duval Hernández 2005).

The conclusion is that at both levels, macro and micro, it makes an important qualitative difference which mobility concept is being gauged. When a layperson asks an economist which of two situations is the more mobile, the answer 'It depends' is not very satisfying. An answer of the type 'Current incomes are more dependent on past incomes in the United Kingdom than in the United States (that is, the UK is *less* mobile in this respect than the USA), but the United Kingdom has more quintile movement than the United States (and therefore is *more* mobile than the USA in this sense)' is more informative, even if less clear-cut than the questioner may have been hoping for.

The Axiomatic Approach to Income Mobility

We have seen that there are different income mobility concepts and that the indices measuring these concepts behave differently from one another. How is the analyst to decide which notion(s) best capture(s) the essence of 'income mobility' for him or her? One approach is to proceed axiomatically, that is, to say that 'for me, mobility is such and such' and then to see which concepts, if any, embody these axioms.

Two broad approaches to axiomatization may be found in the literature. In one approach, mobility is conceptualized in social welfare terms (Atkinson 1980; King 1983; Chakravarty et al. 1985; Dardanoni 1993; Gottschalk and Spolaore 2002; Ruiz-Castillo 2004). In the other, a descriptive approach is used, wherein analysts

specify the properties they wish income mobility concepts and measures to possess, and then proceed to deduce which indices, if any, have these properties (Cowell 1985; Fields and Ok 1996, 1999b; D'Agostino and Dardanoni 2005). The work of Shorrocks (1978a, b) makes use of both of these approaches. This difference between the ethical and the descriptive axiomatizations in the mobility literature parallels the two strands of the inequality literature (Foster and Sen 1997): for Atkinson (1970), inequality is the amount of social welfare lost because incomes are distributed the way they are rather than being distributed perfectly equally, whereas for Sen (1973, p. 2), inequality is objective in the sense that 'one can distinguish between (*a*) "seeing" more or less inequality, and (*b*) "valuing" it more or less in ethical terms'. Note that under both the ethical and the descriptive approaches the amount of mobility recorded has or may have welfare significance. For example, many observers would say that an economy with more directional income movement has performed better than an economy with less directional income movement.

The literature offers a wide variety of axioms, some of which were designed with particular mobility concepts in mind, others of which have been explored to help sharpen what is meant by 'mobility'. Shorrocks (1993) presents 12 axioms for mobility and shows that they are mutually incompatible. In view of their incompatibility, there is a need for judgements as to which ones an analyst wants a measure to embody.

Fields and Ok (1999a) and Fields (2001) have suggested that analysts choose among the axioms by considering their views on simple examples. For example, consider the following three situations:

$$\begin{aligned} \text{I} &: (1,3) \rightarrow (1,3) \\ \text{II} &: (1,3) \rightarrow (2,6) \\ \text{III} &: (2,6) \rightarrow (4,12) \end{aligned}$$

and the corresponding degree of mobility $m(x, y)$. (As above, $\rightarrow$ denotes a change in the ordered (personalized) vector of incomes.) The axiom of strong relativity, if accepted, would maintain that $m(\lambda x, \alpha y) = m(x, y)$ for all $\lambda, \alpha > 0$ and all $x, y \in \mathcal{R}^n_+$. If strong relativity is accepted, it requires that Situations I, II, and III all have the same mobility. In Situation I, the only sensible amount of mobility for there to be is zero, and therefore strong relativity requires that Situations II and III also have zero mobility. An analyst who sees non-zero income mobility in Situations II and III is therefore not a strong relativity adherent.

Similarly, (weak) relativity specifies that $m(\lambda x, \lambda y) = m(x, y)$ for all $\lambda > 0$ and all $x, y \in \mathcal{R}^n_+$. This axiom requires that Situations II and III have the same mobility, though not necessarily the same mobility as Situation I. Therefore, an analyst who sees more mobility in Situation III than in Situation II is not a (weak) relativity adherent either.

The literature offers characterizations of some of the mobility measures that have been used – for example, Fields and Ok's (1996, 1999b) measures of non-directional and directional income movement and Chakravarty et al. (1985) index of mobility as welfare change. More commonly, though, the axioms are used to state a number of desirable properties and then display a measure or a family of measures consistent with these properties.

In summary, a fruitful way for the analyst to choose which mobility concept(s) is (are) most salient for oneself is to consider the axiomatic judgements underlying each of the concepts. To date, some but not all of the income mobility concepts have been so characterized.

Other Issues

The income mobility literature has a number of other issues that remain more or less contentious, not because the different views have not been worked out but because different analysts hold genuinely different positions on a number of important matters.

Is All Distributional Change 'Mobility' or Only Some of It?

Lurking in the background of some writings on income mobility is a fundamental difference of opinion about what income mobility is. For the

majority of analysts, the notion of 'income mobility' has both absolute and relative components. For example, if all incomes double, most would judge there to be more mobility than if all incomes remain unchanged. For some analysts, though, the notion of 'income mobility' is relative only; therefore, the change in the mean needs to be taken out, and 'mobility' applies only to what is left.

Thinking of 'mobility' in this way can lead to some controversial judgements. For example, Chakravarty, Dutta and Weymark (hereafter CDW) (1985) propose the following mobility index:

$$M_{CDW} \equiv \left(E\left(y_{agg}\right)/E(b)\right) - 1,$$

where $E(.)$ is an equality measure, y_{agg} is a vector of aggregate incomes over the observation period, and b is the benchmark vector of incomes under the assumption of complete relative immobility following the first period. In the case in which $E(.)$ is a relative equality measure, the term $E(b)$ is replaced by $E(x)$, where x is the vector of first-period incomes. In the view of these authors (CDW, 1985, p. 8): 'Socially desirable mobility is associated with income structures having positive index values while socially undesirable mobility is associated with income structures having negative index values.' Thus, given their index, CDW judge that mobility contributes positively to social welfare if and only if y_{agg} is distributed more equally than x. Thus, if all incomes rise but the percentage gains are larger at the top end of the income distribution than they are at the bottom, mobility would be judged by CDW to have been socially *undesirable*, in direct contradiction to the quasi-Paretian welfare judgement that an increase in some incomes with no decline in others *raises* social welfare. This difference of views – whether 'income mobility' includes the growth aspect of distributional change or whether 'mobility' is what remains after growth has been taken out – underlies much of the mobility literature, but rarely is it made explicit.

What Is 'Relative Mobility'?

As already noted, the term 'relative mobility' is used ambiguously, sometimes to refer to mobility notions characterized by strong relativity $m(\lambda x; \alpha y) = m(x; y)$ for all λ; $\alpha > 0$ and all $x, y \in \mathcal{R}^n_+$ and sometimes to refer to those characterized by weak relativity $m(\lambda x, \lambda y) = m(x, y)$ for all $\lambda > 0$ and all $x, y \in \mathcal{R}^n_+$. Note that for both of these relativity notions the basis for determining whether a given individual is experiencing upward or downward relative mobility is that individual's change in *income* relative to the *income* changes of others.

However, the term 'relative mobility' is used in yet another sense, namely, to refer to *positional* movements. On this view, an individual experiences relative mobility if and only if he or she changes position (quintile, decile, centile, or rank) from base year to final year. For example, Jenkins and Van Kerm (2003) break down trends in income inequality into a 'pro-poor income growth' component and an 'income mobility' component. The 'income mobility' component involves re-rankings and only re-rankings. Thus, for them as for some others, mobility *is* positional movement and nothing more.

Finally, D'Agostino and Dardanoni (2005) have yet a different definition of relative mobility. For them, relative mobility involves a change in an individual's relative standing with respect to all others, whereas absolute status is something that can be derived by looking at data regarding the individual taken in isolation.

This last point raises the issue of what is meant by 'absolute mobility,' to which we now turn.

What Is 'Absolute Mobility'?

The term 'absolute mobility' is used in at least three different ways in the income mobility literature. One way is to express a concern with gains and losses of *income* rather than *income shares* or *positions*. In this sense, the concept of directional income movement and the various measures of that concept are about absolute mobility. Second, 'absolute mobility' is sometimes used to mean that the analyst is concerned with the *absolute value* of income changes, as would be the case

in studies of non-directional income movement, or flux. Third, the term is used in the sense of *translation invariance*, in the sense that, if all initial and final incomes are increased by the same amount, the new situation has the same absolute mobility as the original one, that is, $m(x + \alpha, y + \alpha) = m(x, y)$.

As is the case elsewhere in economics, when a term has more than one meaning within the same literature, it is probably best to drop the term altogether. Henceforth, researchers would do better to speak of dollar-based, absolute-value-based, or translation-invariant income mobility measures in preference to 'absolute mobility'.

Is 'Income Mobility' Decomposable, and If So, How?

Consider the total income mobility recorded in a population. Under what circumstances can the total be broken down into component parts?

Of the six income mobility concepts considered above, one involves the time-independence aspect of mobility and the other five involve the movement aspect of mobility. The time-independence aspect of mobility is not decomposable. However, there have been decompositions of various movement measures.

One type of decomposition is subgroup decomposability, that is, if the population is divided into J subgroups, the total income mobility in the population as a whole equals a (possibly) weighted average of the mobility in each of the subgroups:

$$m(x, y) = \sum_{j=1}^{J} w_j m_j(x, y).$$

A number of income mobility measures are subgroup decomposable; examples are Fields and Ok's (1996, 1999b) non-directional income movement measures

$$m_1(x, y) \equiv \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|$$

$$\text{and } m_2(x, y) \equiv \frac{1}{n} \sum_{i=1}^{n} |\log y_i - \log x_i|$$

and their directional income movement measure

$$m_3(x, y) \equiv \frac{1}{n} \sum_{i=1}^{n} (\log y_i - \log x_i).$$

A second kind of decomposition is into substantively meaningful components. There is a long tradition in the sociology literature (for example, Bartholomew 1982) of breaking down the movement of individuals among occupations or social classes into two component parts: (*a*) changes that can be attributed to the increased availability of positions in the better occupations and social classes ('structural mobility') and (*b*) changes that can be attributed to increased movement of individuals among occupations and social classes for a given distribution of positions among these classes ('exchange mobility'). Bridging the economics and sociology literatures, Markandya (1982, 1984) proposes two alternative decompositions of income mobility along these lines. The first defines exchange mobility as the proportion of the change in welfare that could have been obtained if the income distribution had stayed constant through time, in which case structural mobility is defined as the residual welfare change. The second defines structural mobility as the change in welfare that would have taken place if the two-period or two-generation transition matrix had exhibited complete immobility, in which case exchange mobility is defined as the residual. Along similar lines, Ruiz-Castillo (2004) shows how the CDW (1985) index of welfare due to mobility could be decomposed into either (*a*) a precisely defined structural component and a residual representing exchange mobility or (*b*) a precisely defined exchange component and a residual representing structural mobility. In all these cases, the residual component makes the decomposition exact but in a rather unexciting way.

The results just cited do *not* mean that an exact additive decomposition of income mobility is impossible. Fields and Ok (1996) show that their mobility index $m_1(x, y)_n \equiv \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|$ is decomposable into the sum of appropriately

defined structural and exchange components. In the case of a growing economy, the decomposition equation is $m_1(x,y) = \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\right) + 2\sum_{\{i:y<x_i\}}(x_i - y_i)$. An analogous decomposition holds for a contracting economy. Along similar lines, Fields and Ok (1999b) show that their directional movement measure $m_3(x,y)_n \equiv \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log x_i)$ is decomposable into social utility growth and social utility transfer components. In all of these cases, the weakness of Markandya's and Ruíz-Castillo's residual approaches is averted.

What Other Empirical Issues Arise?

Empirical researchers should bear in mind two additional issues. One is that, as an empirical matter, the longer the observation period, the greater is the amount of mobility registered (Atkinson et al. 1992). Therefore, care should be taken not to compare, for example, two-year mobility in one context with, for example, five-year mobility in another.

Second, measurement error is a serious issue. There is an ample literature on mismeasurement of earnings *levels* but, as yet, only a very limited literature on mismeasurement of earnings *changes* (Deaton 1997; Bound et al. 2001). A task for the future is to estimate empirically the effect of measurement error on estimates of both macro-mobility and micro-mobility.

Conclusions

The income mobility literature is fundamentally unsettled. This is because the very term 'income mobility' connotes different things to different people. This article has reviewed a number of dimensions in which differences arise: which of six notions most accurately captures the fundamental idea of 'income mobility', which indices best measure each of the concepts, which axioms best characterize the essence of 'income mobility', how income mobility has been evolving over time in different countries, which demographic groups have more mobility than others in different settings, and which theoretical refinements to the notion of 'income mobility' hold the greatest promise.

Given the unsettled state of the field, before researchers 'do a mobility study', it is important that we specify which concept or concepts of mobility we are considering, which measures of these concepts we are using, and which questions we are answering. More than once, when I have given seminars, a member of the audience has raised his or her hand and said, 'But that's not what mobility *is*'. Let us do all that we can to clarify what we are talking about so that we do not talk past one another any more than we have to.

See Also

- ▶ Inequality (Measurement))
- ▶ Intergenerational Income Mobility
- ▶ Longitudinal Data Analysis

Bibliography

Atkinson, A. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Atkinson, A. 1980. The measurement of economic mobility. In *Inkomensverdeling en Openbare Financién*, ed. P. Eijgelshoen and L. van Gemerden. Utrecht: Het Spectrum.

Atkinson, A.B.., F. Bourguignon, and C. Morrisson. 1992. *Empirical studies of earnings mobility.* London: Harwood Academic Publishers.

Bartholomew, D. 1982. *Stochastic models for social processes*. London: Wiley.

Bound, J., C. Brown, and N. Mathiowetz. 2001. Measurement error in survey data. In *Handbook of econometrics*, ed. J. Heckman and E. Leamer, Vol. 5. Amsterdam: North-Holland.

Buchinsky, M., G. Fields, D. Fougère, and F. Kramarz. 2004. *Francs or ranks? Earnings mobility in France, 1967–1999*. INSEE: Mimeo. Paris.

Chronic Poverty Research Centre. 2004. *The chronic poverty report 2004–05*. Manchester: Chronic Poverty Research Centre.

Chakravarty, S., B. Dutta, and J. Weymark. 1985. Ethical indices of income mobility. *Social Choice and Welfare* 2: 1–21.

Cowell, F. 1985. Measures of distributional change: An axiomatic approach. *Review of Economic Studies* 52: 135–151.

D'Agostino, M. and Dardanoni, V. 2005. *The measurement of mobility: A class of distance indices*. Paper presented to the Society for the Study of Economic Inequality, Palma de Mallorca, Spain, July.
Dardanoni, V. 1993. Measuring social mobility. *Journal of Economic Theory* 61: 372–394.
Deaton, A. 1997. *The analysis of household surveys*. Baltimore: Johns Hopkins University Press.
Duval Hernández, R. 2005. Dynamics of labor market earnings and sector of employment in urban Mexico. Doctoral dissertation, Cornell University.
Fields, G. 2001. *Distribution and development: A new look at the developing world*. Cambridge, MA: MIT Press and Russell Sage Foundation.
Fields, G. 2005. *Does income mobility equalize longer-term incomes? New measures of an old concept*. Mimeo: Cornell University.
Fields, G., and E. Ok. 1996. The meaning and measurement of income mobility. *Journal of Economic Theory* 71: 349–377.
Fields, G., and E. Ok. 1999a. The measurement of income mobility: an introduction to the literature. In *Handbook on income inequality measurement*, ed. J. Silber. Boston: Kluwer.
Fields, G., and E. Ok. 1999b. Measuring movement of incomes. *Economica* 66: 455–472.
Fields, G., J. Leary, and E. Ok. 2002. Stochastic dominance in mobility analysis. *Economics Letters* 75: 333–339.
Foster, J., and A. Sen. 1997. *On economic inequality.* expanded ed. Oxford: Oxford University Press.
Freije, S. 2001. Household income dynamics in Venezuela. Unpublished doctoral. dissertation, Cornell University.
Gottschalk, P., and E. Spolaore. 2002. On the evaluation of economic mobility. *Review of Economic Studies* 69: 191–208.
Jarvis, S., and S. Jenkins. 1998. How much income mobility is there in Britain? *Economic Journal* 108: 1–16.
Jenkins, S. and Van Kerm, P. 2003. Trends in income inequality, pro-poor income growth, and income mobility. Discussion Paper No. 904. Bonn: IZA.
King, M. 1983. An index of inequality, with applications to horizontal equity and social mobility. *Econometrica* 51: 99–115.
Krugman, P. 1992. The rich, the right, and the facts. *American Prospect* 11: 19–31.
Maasoumi, E. 1998. On mobility. In *Handbook of applied economic statistics*, ed. D. Giles and A. Ullah. New York: Marcel Dekker.
Markandya, A. 1982. Intergenerational exchange mobility and economic welfare. *European Economic Review* 17: 301–324.
Markandya, A. 1984. The welfare measurement of changes in economic mobility. *Economica* 51: 457–471.
OECD. 1997. *Employment Outlook 1997*. Paris: OECD.
OECD (Organisation for Economic Co-Operation and Development). 1996. *Employment Outlook 1996*. Paris: OECD.
Ruiz-Castillo, J. 2004. The measurement of structural and exchange mobility. *Journal of Economic Inequality* 2: 219–228.
Sánchez Puerta, M. 2005. *Earnings mobility in urban Argentina*. Unpublished doctoral dissertation, Cornell University.
Schumpeter, J. 1955. *Imperialism and social classes*. New York: Meridian Books.
Sen, A. 1973. *On economic inequality.* New York: Norton.
Shorrocks, A. 1978a. Income inequality and income mobility. *Journal of Economic Theory* 19: 376–393.
Shorrocks, A. 1978b. The measurement of mobility. *Econometrica* 46: 1013–1024.
Shorrocks, A. 1993. On the Hart measure of income mobility. In *Industrial concentration and economic inequality*, ed. M. Casson and J. Creedy. London: Edward Elgar.
Slemrod, J. 1992. Taxation and inequality: a time-exposure perspective. In *Tax policy and the economy*, ed. J. Poterba, Vol. 6. Cambridge, MA: MIT Press for the NBER.
Solon, G. 1999. Intergenerational mobility in the labor market. In *Handbook of Labor Economics*, ed. O. Ashenfelter and D. Card, Vol. 3. Amsterdam: North-Holland.

Income Taxation and Optimal Policies

Louis Kaplow

Abstract

Various economic literatures address the question whether first-best prescriptions for government policy require modification because redistributive income taxation distorts labour supply and cannot achieve the distributive ideal. Perhaps second-best rules for public goods provision, corrective taxation, public sector pricing, and other government activity should reflect concerns about distribution and labour supply distortion. Recent work demonstrates, however, that in basic cases first-best principles remain applicable. Demonstrations make use of income tax adjustments that preserve not only budget balance but also the pre-reform distribution of utility.

Keywords

Commodity taxation; Corrective taxes; Cost–benefit analysis; Distortion; Distribution; Environmental taxation; Externalities; Income taxation; Lump-sum tax; Marginal cost pricing; Optimal taxation; Pigouvian tax; Public goods; Public sector pricing; Ramsey taxation; Redistribution; Regulation; Samuelson, P.; Second best; Taxation

JEL Classifications

D61; D62; D63; H21; H23; H24; H44; H42; Q52; Q58

Optimal policy analysis is complicated by problems of the second best. Two of the most important problems – non-ideal distribution and labour supply distortion – are intimately connected with limitations of income taxation. In a first-best world, individualized lump-sum taxes can be used to achieve any desired distribution without causing distortion. Accordingly, the optimal design of other government policies is dictated by familiar first-best rules: the Samuelson cost-benefit test for public goods, the Pigouvian prescription for externalities to equate the full marginal social costs and benefits, marginal cost pricing for publicly provided goods and services and for regulated utilities, and so forth.

In practice, however, informational limitations require the use of distortionary instruments, notably labour income taxation, so even at the optimum (Mirrlees 1971) the distributive ideal is not achieved. Due to the second-best nature of the optimal income taxation problem, it is natural to consider whether first-best prescriptions for other government policies should be modified in order to assist the redistributive function. In addition, such other policies – most obviously but not exclusively those that raise or expend revenue – may affect labour supply, which also may require modification of standard policy rules. Particularly since the explosion of interest in optimal taxation in the 1970s, extensive literatures have developed to address these issues in each particular context. Much work focuses on distortion, some on distribution, and a portion considers both simultaneously. A range of adjustments to first-best formulas have been proposed, revisions that in general depend on the initially prevailing income tax and on the modification thereof that is assumed to accompany the underlying policy reform.

Another strand of research offers a new view of the second-best problem in each of these areas and allows a substantial synthesis across these seemingly different contexts. To analyse these issues, this literature employs a construction under which the income tax modification hypothesized to accompany any policy change is one that, in combination with the altered policy, holds the distribution of utility constant. In a simple standard model, it turns out that first-best policy principles are applicable without refinements: there is no need for distributive adjustments since distribution is unaffected; and, as it happens, holding distribution constant also leaves labour supply unchanged, rendering unnecessary any adjustments on account of labour supply distortion.

The analysis of income taxation and optimal government policy is best introduced in the most fundamental setting, in which the only question is whether a labour income tax should be supplemented by differential commodity taxes. As will be elaborated in the first section below, the answer is negative in simple cases regardless of whether the initial income tax is optimal, a result that in an important sense displaces principles of Ramsey taxation (and, as will subsequently be noted, other applications of Ramsey principles as well). The next section explains how a range of government policies – including public goods provision, regulation of externalities, and public sector pricing – are all formally analogous to differential commodity taxation. Hence, the results (and qualifications) can readily be extended, which allows for the understanding of second-best problems in these disparate fields to be unified substantially. Two final sections relate the analysis to classical and contemporary work and explore further implications of this approach for second-best policy analysis.

Commodity Taxation

The problem of optimal commodity taxation with labour income taxation can be stated as follows. Individuals choose commodity vectors x and labour effort l to maximize the utility function $u(v(\mathrm{x}), l)$, where v is a subutility function. This form of the utility function entails what is referred to as weak separability of labour: for a given level of after-income-tax income, individuals will allocate their disposable income among commodities in the same manner regardless of the level of labour effort required to earn that level of income.

An individual's budget constraint requires that expenditures, $\boldsymbol{\rho x}(wl)$, not exceed before-tax income, wl, minus income taxes, $T(wl)$, which can be negative, thereby allowing for net transfers; $\boldsymbol{\rho}$ is the consumer price vector, w is an individual's wage, and $\mathrm{x}(wl)$ denotes the consumption vector chosen by an individual who earns wl. Individuals' wages w have density $f(w)$, and the government is assumed to know this density but not each individual's wage, which renders individualized lump-sum taxes infeasible. The consumer price vector $\boldsymbol{\rho}$ is understood as the sum of a producer price vector (taken to be constant and equal to production costs) and a vector of commodity taxes (which, if negative, are subsidies).

The government's maximization problem is to select commodity taxes (equivalently, $\boldsymbol{\rho}$) and an income tax schedule $T(wl)$ to maximize a standard concave social welfare function, subject to meeting a given revenue requirement and to incentive compatibility constraints deriving from individuals' maximization problems. If commodity taxes are taken to be zero, we have the optimal nonlinear income tax problem of Mirrlees (1971).

Atkinson and Stiglitz (1976) demonstrated that, when the income tax is set optimally, commodity taxes should be undifferentiated (i.e., uniform) in this basic setting. The derivation to follow is taken from Kaplow (2006), who does not require that the income tax be optimal and provides a more intuitively accessible approach.

For any commodity tax reform, which changes the consumer price vector from ρ to ρ^*, suppose that the income tax schedule is initially adjusted from $T(wl)$ to $T^{\mathrm{o}}(wl)$ such that $V(\boldsymbol{\rho}^*, T^{\mathrm{o}}, wl) = V(\boldsymbol{\rho}, T, wl)$ for all wl, where V is an indirect subutility function indicating the maximized value of $v(\mathrm{x})$, subject to the budget constraint, where $\boldsymbol{\rho}$, T, and wl are taken as given. That is, one adjusts the income tax schedule to the $T^{\mathrm{o}}(wl)$ that restores the original level of subutility achieved at each level of disposable income; hence, $T^{\mathrm{o}}(wl) - T(wl)$ is the schedule of utility-compensating changes in disposable income.

This income tax schedule adjustment has a number of properties. First, if individuals do not change their level of labour supply, they achieve the same utility, for u depends only on v (which is held fixed, given l) and l.

Second, faced with this income tax adjustment, individuals will not in fact change their level of labour supply: each individual's (each type w's) total utility u for any choice of l after this combined reform of commodity taxes and the income tax precisely equals the total utility for that choice of l before the reform; therefore, whatever l previously maximized utility must continue to do so.

Third, the hypothetical reform will in general affect government revenue. Specifically, it can be shown that there will be a surplus if and only if the reform increases efficiency in the narrow sense – by reducing aggregate distortion among commodities – a condition that will prevail, for example, if all commodity taxes (and subsidies) are moved proportionally toward zero, including the case of complete abolition of differential commodity taxation. The reason is that reducing consumption distortion, *ceteris paribus*, raises individuals' utilities; because the income tax adjustment is set to hold utility constant, it must therefore reduce individuals' disposable income to offset what would otherwise be a utility increase. Accordingly, net tax collections must rise.

Finally, to complete the analysis, budget balance can be restored by further adjusting T to rebate the surplus pro rata: $T^*(wl) = T^{\mathrm{o}}(wl) - c$, where c is some positive constant. The result is a Pareto improvement, for utility was unchanged until this final stage of the reform. To summarize, if any commodity tax reform is

accompanied by an income tax adjustment that, when combined with the underlying reform, holds utility constant (until the rebate stage), there is no effect on distribution, labour supply is unchanged, and there is a surplus, allowing a Pareto improvement, if and only if the underlying commodity tax reform is efficient in a narrow, conventional sense.

It is useful to consider the intuition behind this result. It is familiar from the general theory of second-best analysis (Lipsey and Lancaster 1956) that first-best conditions do not generally govern once some distortion is introduced. However, in the present setting the only unavoidable distortion is of the labour–leisure choice, and differential commodity taxation does not help to alleviate it. Thus, differential taxes involve the cost of distorting consumption without any offsetting benefit. The reason that differential commodity taxes cannot help offset the labour–leisure distortion is the assumption of weak separability. Just as different levels of labour supply do not change preferences among commodities, so different consumption allocations do not change the disutility of labour.

This result on the inefficiency of differential commodity taxation provides an important benchmark for understanding and analysis. The conclusion is subject to many qualifications, each of which is best appreciated by reference to this basic starting point. First, as follows immediately from the preceding remarks, weak separability may be violated. This is the point, first elaborated by Corlett and Hague (1953), that it tends to be efficient to tax leisure complements (perhaps beach attendance or reading) and subsidize complements to labour (possibly central city transit or amenities). Second, preferences were taken to be homogeneous, but if preferences depend on unobservable ability it would be optimal to tax commodities preferred by the more able (independent of income per se), perhaps highbrow art, and to subsidize those preferred by the less able. Additional qualifications have been offered, including, importantly, concerns with administration and tax avoidance that may affect income taxation, especially in developing countries.

The foregoing analysis is usefully contrasted with that of Ramsey (1927) taxation, which involves a substantial, widely known literature that itself provides the foundation for much economic analysis of myriad other policy applications (including all those examined in the following section). Most familiar is the rule that commodity taxes should be inversely proportional to the elasticity of demand, with refinements for demand interdependencies. Also well known are modifications due to distributive concerns, which favour taxing luxuries and subsidizing necessities, commands that often conflict with the inverse elasticity rule and thus require tradeoffs (Feldstein 1972; Diamond 1975). As initially emphasized in Atkinson and Stiglitz (1976), however, neither prescription is apt if there is an income tax. In the original Ramsey model in which all individuals are identical and thus there are no distributive concerns, the optimal tax is a uniform lump-sum extraction (a limiting case of an income tax), which, it should be noted, neither requires information about individuals' types nor is distributively objectionable in this setting. When differences in earning ability are admitted, the optimal tax is a nonlinear income tax, and in typical cases the lump-sum component involves a uniform lump-sum subsidy. Nevertheless, optimal commodity taxation still is not guided either by the familiar inverse-elasticity rule or by the general preference for harsher treatment of luxuries than of necessities; as noted, in the basic case, optimal differentiation is nil regardless of the demand elasticity or how demand changes with income.

Paradoxically, the literatures that build upon Ramsey's path-breaking contribution are motivated by second-best concerns, yet it turns out that a more complete second-best analysis – notably, incorporating the income tax, the primary distributive tool and also a central cause of unavoidable distortion that calls for second-best inquiry – returns us to a simple, first-best rule in the benchmark case. Here, that prescription is against differential commodity taxes on account of the resulting distortion of consumption. As will now be explained, this pattern of analysis is replicated with regard to a broad range of government policies.

Government Policies Generally

The foregoing framework can be employed to address the optimal provision of public goods, the optimal control of externalities, and other government actions, as developed by Kaplow (1996, 2004, 2008). The reason is that departures from firstbest rules in these contexts are formally analogous to differential commodity taxation and hence are inefficient in the basic case (a conclusion that also is subject to similar qualifications).

To see this, suppose now that individuals have the utility function $u(v(\mathrm{x}, \mathrm{e}, \mathrm{g}), l)$. Here, e is a vector of externalities (suppose, for example, that each element of e is the population's total consumption of the corresponding commodity in the vector x), and g is a vector of public goods. This functional form maintains the assumption that labour is weakly separable from other sources of individuals' utility.

We can again consider reforms, here of commodity taxes (and subsidies) $\boldsymbol{\rho}$, but now with the thought of internalizing externalities, or of g. Again, we can construct $T^{\circ}(wl)$ such that individuals' subutility v is kept constant if they choose to supply the same level of labour. As before, this reform is distribution neutral and in fact induces all individuals to supply the same labour effort. (A review of the foregoing analysis will confirm that nothing depended on the fact that the reform was only of commodity taxes or that there were no externalities or public goods involved.)

The question, then, is whether the intermediate adjustment of the income tax schedule, from $T(wl)$ to $T^{\circ}(wl)$, will produce a surplus or a deficit. With externalities, if, for example, one sets all commodity taxes equal to the marginal external effect of consumption on individuals' utilities – the traditional Pigouvian prescription (Pigou 1920) – there will be a surplus: individuals may be better or worse off because of being subject to a different vector of commodity prices, and they may be better or worse off on account of changes in the levels of externalities; however, it can be demonstrated that the net effect on revenue is positive, essentially because of traditional efficiency considerations. (Note that the income tax adjustment from $T(wl)$ to $T^{\circ}(wl)$ taxes away all sources of surplus and compensates for any disutility; hence, the sign of the net revenue effect is given by the sign of the total of all changes in individuals' surplus from the underlying reform.) Observe that this result is very similar in spirit to that on commodity taxation without externalities. There, the optimum involves setting consumer prices equal to true marginal resource costs of commodities; with externalities, the same principle holds, but true resource costs now include not only production costs but also effects on others' utilities.

For public goods, the total revenue effect has two components. The first (which is negative) is the production cost of the public goods, and the second is (by the method of construction of $T^{\circ}(wl)$) the integral of individuals' surplus from changes in the levels of the public goods. Hence, there is a surplus (deficit) if and only if the reform passes (fails) the Samuelson (1954) cost–benefit test, which asks whether the integral of individuals' benefits exceeds the cost of producing the public goods. The essence of the argument is again similar to that for the basic case with commodity taxation. For example, supplying less of a public good than dictated by the Samuelson test corresponds to imposing a differential tax on a private good. To push the analogy further, consider a hypothetically decentralized regime in which consumer prices for private goods correspond to Lindahl prices for public goods, and commodity taxes on public goods are defined as the difference between the price charged to a consumer in the imaginary regime and that consumer's marginal rate of substitution. The source of the allocative inefficiency is again a failure of the prices faced by consumers to equal true marginal resource costs.

In the present setting, therefore, moving to the first best – now regarding internalization of externalities or provision of public goods rather than setting commodity taxes in a simpler world – makes possible a Pareto improvement. Concerns about distribution and labour supply effects caused by the income tax can be ignored because they are moot.

Similar logic can be employed to address other areas of government policy, most obviously regulations that mimic corrective taxation but also seemingly unrelated fields like public sector pricing and utility regulation. Thus, marginal cost pricing will be optimal in spite of distributive concerns or the distortionary cost of raising funds to meet deficits because, if the income tax is adjusted in the manner described, distribution will be unaffected and there will be a net surplus if the reform is (narrowly) efficient in the basic case.

Historical Development of Second-Best Policy Rules

First-best principles have a long and familiar lineage. The command to internalize externalities is inspired by Pigou's (1920) classical treatment, and the cost-benefit test for public goods is due to Samuelson's (1954) elegant formulation. It is notable that Samuelson (1954) explicitly said that he was considering a first-best setting in which individualized lump-sum taxes permitted any social welfare optimum to be implemented.

Second-best qualifications start with another of Pigou's (1928) books, in which he observed that, on account of the resource cost of raising revenue, public goods probably should have to meet a higher standard. Refinements appeared in Atkinson and Stern (1974), Diamond and Mirrlees (1971), and Stiglitz and Dasgupta (1971), with subsequent research crystallized by Ballard and Fullerton (1992). Analogous work on environmental taxation – addressed to the possibility of a 'double dividend' (a tax might both internalize an externality and raise revenue distortion-free) and qualifications implying a more negative view of corrective policies – became intense in the 1990s (see Bovenberg and Goulder 2002; Goulder 2002). Largely separate literatures proposed second-best adjustments to account for distributive effects (Weisbrod 1968; Drèze and Stern 1987). See also Bös (1985) on public sector pricing.

Much of this work builds on Ramsey's (1927) model of taxation and extensions thereof. Often, such analyses employ the original representative-individual model in which distribution is immaterial; yet, at the same time, the possibility of income taxation is ignored (specifically, the possible use of a uniform grant that, as noted above, makes commodity taxation unnecessary) or the income tax adjustments that are stipulated turn out not to be distribution-neutral. Literature focusing on distribution also often ignores the availability of the income tax.

The lessons presented in the prior sections arise from another line of work that developed intermittently and largely independently of the foregoing literatures. Hylland and Zeckhauser (1979) used a distribution-neutral income tax adjustment with a special case of individuals' utility functions to show that distributive weights are inappropriate in cost–benefit analysis. Shavell (1981) offers a similar demonstration for legal rules. Christiansen (1981) and Boadway and Keen (1993) show that, with an optimal income tax, the basic cost–benefit test for public goods is appropriate. Kaplow (1996, 2004, 2006, 2008) considers both distribution and labour supply distortion, does not require the income tax to be optimal, and examines a broad range of government policies.

Implications

Ever since Lipsey and Lancaster (1956), economists have sought to develop principles to provide guidance in a second-best world; indeed, in the area of taxation, the search had already begun. The inability to achieve an ideal distribution without distortion is one of the most important unavoidable deviations from the first best. Thus, not surprisingly, substantial research addresses second-best concerns regarding income taxation and commodity taxation as well as all manner of government policies that may have distributive effects or influence government revenue.

Perhaps surprisingly, a number of first-best principles prove to be rather robust in basic, benchmark cases. Important caveats were noted,

but, importantly, they are largely orthogonal to the original second-best concerns that motivate most research in these fields.

One further qualification deserves attention. The present analysis assumes that the income tax will be adjusted in a distribution-neutral manner. This is hardly an unnatural assumption. For example, if the initial income tax does not optimally trade off distribution and distortion, the divergence may arise from political forces that dictate some other degree of redistribution. If so, particular reforms might be expected to leave that distributive balance unaltered.

Nevertheless, consider the possibility of non-distribution-neutral adjustments of the income tax. As suggested in Kaplow (1996, 2004, 2008), a simple two-step decomposition is illuminating in this case:

1. Assume that, initially, the underlying policy is implemented in the previously hypothesized distribution-neutral fashion.
2. Assume also that, a moment later, a further income tax adjustment transforms the policy in step 1 into the actually imagined policy.

Analysis of step 1 can proceed as before. Step 2, observe, is a purely redistributive reform. Accordingly, the analysis is in the province of optimal income taxation and involves the familiar distribution-distortion trade-off. Significantly, the analysis of step 2 is generic – that is, it is the same regardless of whether step 1 involves changing commodity taxes, one or another regulation, the level of some public good, or indeed nothing at all (a purely redistributive overall reform). For economists, this allows substantial specialization. Step 2 analysis must be undertaken anyway and, as noted, tends to be independent of step 1. Step 1 analysis can be undertaken by experts on gasoline taxes, health care, electric utilities, and so forth, who need not concern themselves with redistribution. Policymakers can combine analyses as appropriate.

Specialization has an additional virtue in this context: it facilitates communication, both among researchers and to policymakers. For example, a study of a highway project that does not focus on step 1 will need to include analysis of (*a*) direct effects of the highway project (such as on pollution or congestion), (*b*) what other, budget-accommodating tax adjustment will in fact be made in the long run (an exercise in political economy), (*c*) an analysis of the effects of the resulting change in the extent of redistribution, and (*d*) a social welfare assessment, requiring choice of a social welfare function. Relatedly, when studies of a highway project reach different conclusions, the discrepancies may arise from any combination of these four components, making it difficult to compare and synthesize research.

A particular concern arises with much work in these literatures, both abstract and highly applied, because step 1 is often combined with an incomplete analysis of step 2. For example, work might identify a redistributive benefit from a policy; yet, if there is not a complete analysis of redistributive taxation, the likely associated increase in labour supply distortion may be overlooked. Contrariwise, much work identifies increases in distortion, failing to recognize that the increases are due to effects on labour supply that accompany an implicit increase in redistribution, the benefit of which is omitted. Because of the original second-best problem, involving redistribution through distortionary taxation, redistribution is not an unambiguous good because (usually) it comes at a cost, and distortion – particularly of labour supply – is not an unmitigated evil because (frequently) it is symptomatic of an underlying benefit. Analysis that incorporates one side of the balance while excluding the other may be the worst approach of all.

To summarize, Ramsey principles are widely acknowledged and broadly employed as a foundation for second-best policy analysis. However, at least in developed economies in which an income tax is feasible, the model's most familiar implications for differential commodity taxation are inapt and, by extension, so are its applications to public goods provision, regulation of externalities, public sector pricing, and other policy areas. In the basic case, the problem of optimal

redistribution – involving the trade-off of distribution and labour supply distortion – is separable from these other realms. Accordingly, traditional first-best principles that focus on efficiency in the area under consideration provide a useful benchmark. Complications abound, but for the most part they do not replicate the adjustments called for by the original Ramsey model or typical applications thereof. Instead, they are best understood by direct reference to the problem of redistributive income taxation.

See Also

- ▶ Compensation Principle
- ▶ Environmental Economics
- ▶ Mirrlees, James (Born 1936)
- ▶ Optimal Taxation
- ▶ Pigouvian Taxes
- ▶ Public Goods
- ▶ Ramsey Model
- ▶ Redistribution of Income and Wealth
- ▶ Taxation of Income

Acknowlgments I am grateful to Steven Shavell for comments and the John M. Olin Center for Law, Economics, and Business at Harvard University for financial support.

Bibliography

Atkinson, A., and N. Stern. 1974. Pigou, taxation, and public goods. *Review of Economic Studies* 41: 119–128.

Atkinson, A., and J. Stiglitz. 1976. The design of tax structure: Direct versus indirect taxation. *Journal of Public Economics* 6: 55–75.

Ballard, C., and D. Fullerton. 1992. Distortionary taxes and the provision of public goods. *Journal of Economic Perspectives* 6: 117–131.

Boadway, R., and M. Keen. 1993. Public goods, self-selection and optimal income taxation. *International Economic Review* 34: 463–478.

Bös, D. 1985. Public sector pricing. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, Vol. 1. Amsterdam: Elsevier.

Bovenberg, A., and L. Goulder. 2002. Environmental taxation and regulation. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, Vol. 3. Amsterdam: Elsevier.

Christiansen, V. 1981. Evaluation of public projects under optimal taxation. *Review of Economic Studies* 48: 447–457.

Corlett, W., and D. Hague. 1953. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21: 21–30.

Diamond, P. 1975. A many-person Ramsey tax rule. *Journal of Public Economics* 4: 283–299.

Diamond, P., and J. Mirrlees. 1971. Optimal taxation and public production. II: Tax rules. *American Economic Review* 61: 261–278.

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.

Feldstein, M. 1972. Equity and efficiency in public sector pricing: The optimal twopart tariff. *Quarterly Journal of Economics* 86: 175–187.

Goulder, L. 2002. *Environmental policy making in economies with prior tax distortions*. Cheltenham: Edward Elgar.

Hylland, A., and R. Zeckhauser. 1979. Distributional objectives should affect taxes but not program choice or design. *Scandinavian Journal of Economics* 81: 264–284.

Kaplow, L. 1996. The optimal supply of public goods and the distortionary cost of taxation. *National Tax Journal* 49: 513–533.

Kaplow, L. 2004. On the (ir)relevance of distribution and labor supply distortion to government policy. *Journal of Economic Perspectives* 18: 59–75.

Kaplow, L. 2006. On the undesirability of commodity taxation even when income taxation is not optimal. *Journal of Public Economics* 90: 1235–1250.

Kaplow, L. 2008. *The theory of taxation and public economics*. Princeton: Princeton University Press.

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

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 68: 175–208.

Pigou, A. 1920. *The economics of welfare*. London: Macmillan.

Pigou, A. 1928. *A study in public finance*. London: Macmillan.

Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Shavell, S. 1981. A note on efficiency vs. distributional equity in legal rulemaking: Should distributional equity matter given optimal income taxation? *American Economic Review* 71: 414–418.

Stiglitz, J., and P. Dasgupta. 1971. Differential taxation, public goods, and economic efficiency. *Review of Economic Studies* 38: 151–174.

Weisbrod, B. 1968. Income redistribution effects and cost–benefit analysis. In *Problems in public expenditure analysis*, ed. S. Chase. Washington, DC: Brookings Institution.

Income–Expenditure Analysis

Michael Artis

The term 'income–expenditure analysis' serves as a short-hand expression for the dominant type of conceptual framework for macroeconomic analysis to emerge from the debate which crystallized around Keynes's *General Theory* (Keynes 1936). As Coddington (1976) notes, income-expenditure analysis was not the only thing to be learned from the *General Theory*, but it has certainly been the dominant one, forming the central message of Keynesian economics as generally understood. Although the term does not appear to have been used by Keynes himself it is to be found, freely used, in the early works of exposition of the *General Theory* and the Keynesian Revolution. At the formal and simplest level it can be taken to refer to the 45° 'Keynesian Cross' diagram, at a more sophisticated level to the IS/LM analysis.

At the outset of the *General Theory*, Keynes noted the inability of traditional theory to explain the Great Depression. His analysis was evolved to make good this deficiency and does so by sidestepping the concerns of that theory, in that the power of the price system to ensure the equilibration of the economy is simply denied. Relative prices being set on one side, the analysis focuses on the interaction of flows of expenditure to explain economic fluctuations and the determination of output. In the more familiar embodiments of income-expenditure analysis, the abstraction is clearer and proceeds further than it does in the *General Theory* itself. In particular, both wages and prices are taken as fixed, whilst the stock of productive capital, wealth and the 'state of expectations' are also taken as given. Capital markets are imperfect. Keynes's concept of the propensity to consume is central and leads to the key result of the multiplier. Because expenditures are cash-constrained and output is demand-constrained, an autonomous increment in demand relaxes the constraint on spending and so on output, raising incomes and so stimulating further 'rounds' of expenditures. The ultimate increase in demand may exceed the initial stimulus. In the simplest version of this analysis, exemplified in the 'Keynesian Cross' diagram, asset prices are taken as fixed along with wages and commodity prices. In the IS/LM version this restriction is relaxed and demand increases may then spend themselves at least partly in changes in asset prices which in turn affect the desire to acquire capital goods. In this way the multiplier process may be attenuated if monetary conditions are tight rather than passive.

The framework readily accommodates policy, inviting the tools of fiscal and monetary policy to be used in the management of demand with the objective of output stabilization. Further, the analysis lends itself to quantification; with the addition of a modelling of the wage-price sector and of the foreign exchanges (or capital flows), income–expenditure analysis was the initial basis for the construction of macroeconometric models of the kind used by Finance Ministries and Central Banks and indeed remains the basic foundation for such models today.

The abstraction from relative prices of this approach, its highly aggregative and often heavily quantified nature has exposed income-expenditure analysis to the criticism that it is excessively mechanical (Coddington 1976) describes it as 'hydraulic' Keynesianism in reference to this criticism and perhaps to the fact that some early teaching machines of this model were literally hydraulic); it was said to 'lack microfoundations'. Modern temporary equilibrium theory of the kind pioneered by, for example, Barro and Grossman (1976) has supplied such foundations – or, to be more accurate, has shown how a general equilibrium model with fixed prices generates properties which are very similar to those to be found in income–expenditure analysis – involuntary unemployment may exist, the multiplier process is replicated, and fiscal and monetary policy can be given a demand management rationale, for example. The condition on which these results are generated is that wages and prices fail to equilibrate the model and this condition is imposed as a stylized fact rather than explained by the model itself, a weakness in the

eyes of critics who would argue that the existence of such rigidities implies unexploited opportunities for profitable trade to (well-informed) rational agents. On the other hand, there appear to be a number of reasons why wages and prices are in fact sticky, supporting the principal strategic simplification of income-expenditure analysis, whilst modern exploration of the properties of rational expectations models has confirmed the wisdom of Keynes's tactic of treating expectations as parametric in view of the problem of multiple equilibria. Certainly, the resultant mode of analysis has been very successful and still retains a dominant place in macroeconomics, despite the flourishing new classical school.

See Also

- ▶ Autonomous Expenditures
- ▶ Inflationary Gap
- ▶ IS–LM Analysis
- ▶ Neoclassical Synthesis

Bibliography

Barro, R., and H.I. Grossman. 1976. *Money, employment and inflation*. London/New York: Cambridge University Press.

Coddington, A. 1976. Keynesian economics: The search for first principles? *Journal of Economic Literature*, December. Reprinted in A. Coddington, *Keynesian economics: The search for first principles*. London: George Allen & Unwin, 1983.

Keynes, J. M. 1936. *The general theory of employment, interest and money*. Reprinted in *The collected writings of John Maynard Keynes*, vol. VII. London: Macmillan, 1971.

Incomes Policies

Rupert Pennant-Rea

Keynes's *General Theory* attacked the foundation of the quantity theory of money – the proposition that the level of activity is determined by real forces. But though Keynes provided a new theory of output, he offered no systematic explanation of the price level. He simply took money wages as given, and argued that their level was the key determinant of all nominal magnitudes.

In *How to Pay for the War* Keynes took the analysis further, presenting two complementary theories of the determination of the *rate of change* of money wages and prices. The rate of inflation was affected on the one hand, by the pressure of demand; on the other, by the attempt of workers to maintain the real value of their incomes in a recession or when the terms of trade deteriorated.

These two explanations reinforce one another in Keynes's exposition, but they are formally distinct. The pressure of demand may explain the rate of change of wages independently of any predetermined real wage; this was one central implication of Phillips's (1958) study of the relationship between money wage inflation and the level of unemployment, and of later refinements of the Phillips curve. However, the efforts of workers to maintain real incomes may determine the rate of change of money wages and prices relatively *independently* of the pressure of demand (though they may be affected by rapid changes in the pressure of demand).

If the two processes are combined, the result may be formulated as a relationship between the pressure of demand and the rate of change of the rate of inflation. In some writers' view, the pressure of demand leads to real wage bids which cannot be satisfied and hence to cumulative rises in wages and prices, as workers and employers bid for shares of real income which total more than one (Rowthorn 1977).

When the Phillips curve analysis was presented, it appeared to carry an important message for macroeconomic policy: that there is a trade-off between unemployment and the rate of inflation (or between unemployment and the rate of change of the rate of inflation). The inflation which was believed to be associated with macroeconomic expansion provided a constraint on such expansion. If inflationary pressures could be reduced, it was argued, economies could produce more. From that perspective, the purpose of an incomes

policy was to reduce the inflationary pressure associated with any given level of demand.

Discussion of incomes policy often fails to make it clear whether the policy is intended to operate solely on nominal wages, or on real wages. In fact, the distinction is important for both analytical and practical reasons. If the level of money wages is given and unrelated to real variables (as Keynes appeared to suggest in the *General Theory*), then it may be argued that the role of incomes policy is to moderate the rate of change of nominal magnitudes, with no particular implications for real wages, or for the distribution of income between wages and profits. In which case, the pressure to disrupt an income policy will logically come from groups intent on changing the distribution of real income in their favour.

But if the level of money wages is the outcome of bargaining over real incomes, then the purpose of an incomes policy will usually be to persuade workers to achieve lower real wages and thus change the distribution of income. In which case, the policy must either deal directly with the forces which determine real incomes in the first place; or it will be placed under considerable strain – perhaps breaking down as those forces reassert themselves (Tarling and Wilkinson 1977).

In reality, incomes policies have often been justified as a way of reducing inflation in nominal magnitudes; but they have also changed the distribution of income, usually by squeezing real wages.

In the view of those who favour incomes policy, successful policies – those that have lasted the longest and been associated with relatively low rates of inflation – have been those which (a) recognize explicitly that the distribution of real income is at stake, and plan the rate of change of money wages as part of a socio-political 'deal' (war-time policies are good examples); and/or (b) are implemented when real national income is growing rapidly, so that all real wages can increase even as the distribution of real income is changed.

Although labelled 'incomes policy', most such policies are concerned only with wages. They may sometimes be linked to controls on dividends to provide an aura of 'fairness'; or to a prices policy, in which case they are overtly a policy for real incomes.

In the OECD countries five main types of incomes policies have been tried. One version relies on *exhortatory guidelines*. It does little more than encourage employers and employees to settle for lower increases in nominal wages and salaries. As one author sympathetic to incomes policy has argued, 'a principal objective of incomes policy must be to inform public opinion and develop a consensus on the appropriate rate of increase for most wages and salaries . . .' (Braun 1986). To this end, a government may set an example itself in the sectors in which it is the direct employer (such as the civil service) or the indirect financier (as may be the case in some nationalized industries). And the government may also argue that monetary and fiscal expansion will be restrained if inflation is 'too high'.

If exhortation fails, a government may resort to *temporary measures*, such as a wage freeze. For example, it may be argued that a freeze is a temporary measure designed to adjust inflationary expectations in such a way as to minimize the consequences of deflationary policies. The most dramatic examples of the use of temporary measures have occurred in countries suffering from very rapid inflation (Argentina, Bolivia and Israel in 1985, and Brazil in 1986), and have often been accompanied by price freezes and currency reform.

A *statutory norm* involves a government laying down limits for the increase of nominal wages allowed in any one year. The increase has usually been couched in percentage terms, though sometimes this has been combined with an absolute limit. The rate is set for the economy as a whole, with increases beyond the norm being sanctioned by some form of arbitration tribunal – usually on the grounds of exceptional productivity growth. This approach has been widely used in Britain, and, from time to time, in the United States. The relative success or failure of the statutory norm has been determined by its consequences for the rate of increase of real wages, and by the impact on wage differentials of productivity-based 'exceptions'.

Recognition that incomes policies are concerned with the distribution of real income

has led to the suggestion that nominal magnitudes should be *indexed*. This usually means that wage agreements contain automatic escalator clauses. Such agreements are undermined when the increase in real incomes implicit in nominal wage negotiations cannot be sustained – as, for example, when the quadrupling of oil prices in 1973–4 resulted in a transfer of real income from the OECD countries to the oil producers. In such circumstances, indexation can become a source of explosive inflation, at least in the short term.

Milton Friedman has claimed (1974) that

> widespread escalator clauses would make it easier for the public to recognize changes in the rate of inflation, would thereby reduce the time-lag in adapting to such changes, and thus make the nominal price level more sensitive and variable . . ..But, if so, the real variables would be less sensitive and more stable – a highly beneficial trade-off.

In short, indexation would allow governments to disinflate (by appropriate monetary measures) with the least harmful consequences for the real economy. Wage earners would find that their wages adjusted quickly to disinflationary policies. Without that prompt adjustment, they would have obtained unjustifiably large increases, which would cause bankruptcies and unemployment. Indexation may be favoured on other grounds. For example, it may give confidence to particular bargaining groups, who will therefore have no incentive to bid for large increases in nominal wages as a means of protecting real wages.

As a policy device, indexation was adopted by several countries in the late 1940s, among them Belgium, Luxembourg, Italy, Denmark and Norway. During the 1970s indexation was introduced in Britain, the Netherlands, Ireland, Switzerland and Australia. Indexation had been common in France until 1958, when it was abolished at the same time as the franc was devalued. This proved a pointer to the subsequent abandonment of indexation in many other countries. Governments found that indexation prevented them from making desired changes in relative prices (such as changes in exchange rates) which had implications for the distribution of income. In such circumstances indexation is a device for institutionalizing inflation.

A different automatic device for regulating wage increases is the *tax-based incomes policy* (TIP).

In this scheme the government sets a norm for wage increases. Firms that pay less than the norm receive a reward, typically in the form of lower corporate taxes. Those that pay more face a tax penalty. A variant would provide employees with tax incentives to settle for wage increases below the norm.

Schemes of this type have been tried only in the most general form. In Britain in 1977–8, for example, the Labour government promised to reduce income tax if the national increase in wages was moderated. In Austria in 1967–8, the government achieved a wage-tax bargain. When it tried again, in 1974, its proposal was rejected by the unions.

In practice, incomes policies have tended to follow a sequence of initial acceptance and effectiveness, followed by growing opposition and circumvention, and then breakdown. Explanations of this phenomenon vary according to each author's view of the price mechanism. For example, some economists argue that incomes policies fail because they seek to over-ride market forces. An incomes policy unsupported by suitably anti-inflationary macroeconomic policies is bound to fail; but one that is so supported is superfluous. Indeed it may do microeconomic harm because it slows down and distorts market adjustments: 'the theory of incomes policy, as opposed to the desperate "ad-hocery" of practice, has not come to grips with resolving some form of wage, dividend and price control with the resource-allocating function that both goods and factor prices are held to play' (Ball and Doyle 1969; see also Paish 1986).

A different explanation of the failure of incomes policy comes from those who take a somewhat jaundiced view of the efficiency of markets, particularly of the labour market. They attribute failure to inefficient implementation – imposing an incomes policy as part of a deflationary package, rather than using it as part of an expansionary programme. Other analysts point to particular groups of workers who break norms. Still others single out the impact of external shocks, which destroy the assumptions on which the policy had been framed.

See Also

- ▶ Demand Management
- ▶ Full Employment
- ▶ Inflation
- ▶ Stabilization Policy
- ▶ Wage Indexation

Bibliography

Ball, R.J., and P. Doyle (eds.). 1969. *Inflation*. London: Penguin.

Braun, A.R. 1986. *Wage determination and incomes policy in open economies*. Washington: International Monetary Fund.

Friedman, M. 1974. Monetary correction. In *Essays on inflation and indexation*, ed. M. Friedman. Washington, DC: American Enterprise Institute for Public Policy Research.

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

Keynes, J.M. 1940. *How to pay for the war*. London: Macmillan.

Paish, F.W. 1986. The limits of incomes policies. In *Policy for incomes*, Hobart Paper, vol. 29, 4th ed, ed. F.W. Paish and J. Hennessy. London: Institute of Economic Affairs.

Phillips, A.W.H. 1958. The relation between unemployment and money wage rates in the United Kingdom, 1861–1957. *Economica* 25: 283–299.

Rowthorn, R. 1977. Conflict, inflation and money. *Cambridge Journal of Economics* 1: 215–239.

Tarling, R., and S.F. Wilkinson. 1977. The social contract – Post-war incomes policies and their inflationary impact. *Cambridge Journal of Economics* 1(4): 395–414.

Incomplete Contracts

Oliver Hart

The past decade has witnessed a growing interest in contract theories of various kinds. This development is partly a reaction to our rather thorough understanding of the standard theory of perfect competition under complete markets, but more importantly to the resulting realization that this paradigm is insufficient to accommodate a number of important economic phenomena. Studying in more detail the process of contracting – particularly its hazards and imperfections – is a natural way to enrich and amend the idealized competitive model in an attempt to fit the evidence better.

In one sense, contracts provide the foundation for a large part of economic analysis. Any trade – as a quid pro quo – must be mediated by some form of contract, whether it be explicit or implicit. In the case of spot trades, however, where the two sides of the transaction occur almost simultaneously, the contractual element is usually down-played, presumably because it is regarded as trivial (although we will argue below that this need not be the case). In recent years, economists have become much more interested in long-term relationships where a considerable amount of time may elapse between the quid and the quo. In these circumstances, a contract becomes an essential part of the trading relationship.

Research on contracts has progressed along several different lines. Two prominent areas of work are principal-agent theory and implicit labour contract theory. In these literatures, the focus is on risk-sharing or income-smoothing as the motivation for a contract; that is, on the gains the parties receive from transferring income from one state of the world or one period to another. For example, in implicit contract theory, it is supposed that workers are constrained in their ability to get insurance or to borrow on the open market and that employers therefore offer these services as part of an employment contract.

While 'income-smoothing' is undoubtedly important, there are arguably more fundamental factors underlying the existence of long-term contracts. A basic reason for long-term relationships is the existence of investments which are to some extent party specific; that is, once made, they have a higher value inside the relationship than outside. Given this 'lock-in' effect, each party will have some monopoly power *ex-post*, although there may be plenty of competition *ex-ante*, before investments are sunk. Since the parties cannot rely on the market once their relationship is underway, a long-term contract is an important way for them to regulate, and divide up

the gains from, their trade. This will be the case even if the parties are risk neutral and have access to perfect capital markets, that is, even if the income-smoothing role is completely inessential. Moreover, in the case, say, of supply contracts involving large firms, risk neutrality and perfect capital markets may be reasonable approximations in view of the many outside insurance and borrowing/lending opportunities available to such parties.

In spite of their importance, contracts whose *raison d'être* is the regulation of specific relationships have been the subject of little analysis. A notable early reference is Becker's (1964) analysis of worker training. More recently, Williamson (1985) and Klein et al. (1978) have emphasized the difficulty of writing contracts which induce efficient relationship-specific investments as an important factor in explaining vertical integration.

In this entry I will try to summarize what is known theoretically about contracts of this type. I will focus particularly on the problems which arise when the parties write a contract which is incomplete in some respects. Given the rudimentary state of our knowledge of the area, the entry is inevitably quite speculative in nature. The reader who is interested in an elaboration of some of the ideas presented here, and how they fit into the rest of contract theory, might want to consult Hart and Holmstrom (1987).

The Benefits of Writing Long-Term Contracts Given Relationship-Specific Investments

The role of a long-term contract when there are relationship-specific investments can be seen from the following example (based on Grout 1984). Let B, S be, respectively, the buyer and seller of (one unit of) an input. Suppose that in order to realize the benefits of the input, B must make an investment, a, which is specific to S; for example, B might have to build a plant next to S. Assume that there are just two periods; the investment is made at date 0, while the input is supplied and the benefits are received at date 1. S's supply cost at date 1 is c, while B's benefit function is $b(a)$ (all costs and benefits are measured in date 1 dollars).

If no long-term contract is written at date 0, the parties will determine the terms of trade from scratch at date 1. If we assume that neither party has alternative trading partners at date 1, there is, given B's sunk investment cost a, a surplus of $b(a) - c$ to be divided up. A simple assumption to make is that the parties split this 50:50 (this is the Nash bargaining solution). That is, the input price p will satisfy $b(a) - p = p - c$. This means that the buyer's overall payoff, net of his investment cost, is

$$b(a) - p - a = \frac{b(a) - c}{2} - a \qquad (1)$$

The buyer, anticipating this payoff, will choose a to maximize (1), i.e. to maximize $1/2\ b(a) - a$.

This is to be contrasted with the efficient outcome, where a is chosen to maximize total surplus, $b(a) - c - a$. Maximizing (1) will lead to underinvestment; in fact, in extreme cases, a will equal zero and trade will not occur at all. The inefficiency arises because the buyer does not receive the full return from his investment – some of this return is appropriated by the seller in the date 1 bargaining. Note that an upfront payment from S to B at date 0 (to compensate for the share of the surplus S will later receive) will not help here, since it will only change B's objective function by a constant (it is like a lump-sum transfer). That is, it redistributes income without affecting real decisions.

Efficiency can be achieved if a long-term contract is written at date 0 specifying the input price p^* in advance. Then B will maximize $b(a) - p^* - a$, yielding the efficient investment level, a^* An alternative method is to specify that the buyer must choose $a = a^*$ (if not he pays large damages to S) – the choice of p can then be left until date I, with an upfront payment by S being used to compensate B for his investment. The second method presupposes that investment decisions are publicly observable, and so in practice may be more complicated than the first (see below).

We see then that a long-term contract can be useful in encouraging relationship- specific investments. The word 'investment' should be interpreted broadly here; the same factors will apply whenever one party is forced to pass up an opportunity as a result of a relationship with another party (e.g., A's 'investment' in the relationship with B may be not to lock into C). That is, the crucial element is a sunk cost (direct or opportunity) of some sort (an effort decision is one example of a sunk cost). Note that the income-transfer motive for a long-term contract is completely absent here; there is no uncertainty and everything is in present value terms.

Given the advantages of long-term contracts in specific relationships, the question that obviously arises is why we do not see more of them, and why those we do see seem often to be limited in scope. To this question we now turn.

The Costs of Writing Long-Term Contracts

Contract theory is sometimes dismissed because 'we don't see the long-term contingent contracts that the theory predicts'. In fact, there is no shortage of complex long-term contracts in the world. Joskow (1985), for example, in his recent study of transactions between electricity generating plants and mine-mouth coal suppliers finds that some contracts between the parties extend for fifty years, and a large majority for over ten years. The contractual terms include quality provisions, formulae linking coal prices to costs and prices of substitutes, indexation clauses, and so on. The contracts are both complicated and sophisticated. Similar findings are contained in Goldberg and Erickson's (1982) study of petroleum coke.

At a much more basic level, a typical contract for personal insurance, with its many conditions and exemption clauses, is not exactly a simple document. Nor for that matter is a typical house rental agreement. On the other hand, labour contracts are often surprisingly rudimentary, at least in certain respects (for example, there is little indexation of wages to retail prices or to firm employment or sales; layoff pay is limited, etc.).

Given that complex long-term contracts are found in some situations but not others, it is natural to explain any observed contract as an outcome of an optimization process in which the relative benefits and costs of additional length and complexity are traded off at the margin. In the last section, we indicated some of the benefits of a long-term contract. (The example considered was sufficiently straightforward that the ideal long-term contract was a simple noncontingent one; however, with the inclusion of such factors as uncertainty about payoffs and variable quality of the input, the optimal contract would be a (possibly much more complex) contingent one.) But what about the costs? These are much harder to pin down since they fall under the general heading of 'transaction costs', a notoriously vague and slippery category. Of these, the following seem to be important: (1) the cost to each party of anticipating the various eventualities that may occur during the life of the relationship: (2) the cost of deciding, and reaching an agreement about, how to deal with such eventualities; (3) the cost of writing the contract in a sufficiently clear and unambiguous way that the terms of the contract can be enforced; and (4) the legal cost of enforcement.

One point to note is that *all* these costs are present also in the case of *short- term* contracts, although presumably they are usually smaller. In particular, since the short-term future is more predictable, the first cost is likely to be much reduced, and so possibly is the third. However, it certainly is not the case that there is a sharp division between short-term contracts and long-term contracts, with, as is sometimes supposed, the former being costless and the latter being infinitely costly.

It is also worth emphasizing that, when we talk about the cost of a long-term contract, we are presumably referring to the cost of a 'good' long-term contract. There is rarely significant cost or difficulty in writing *some* long-term contract. For example, the parties to an input supply contract could agree on a fixed price and level of supply for the next fifty years. They do not presumably because such a rigid arrangement would be very inefficient. (In some cases the courts will not enforce such an agreement, taking the point of

view that the parties could not really have intended it to apply unchanged for such a long time. A clause to the effect that the parties really do mean what they say should be enough to overcome this difficulty, however. In other cases, it may be impossible to write a binding long-term contract because the identities of some of the parties involved may change. For example, one party may be a government that is in office for a fixed period, and it may be impossible for it to bind its successors. This latter idea underlies the work of Kydland and Prescott (1977) and Freixas et al. (1985)).

Due to the presence of transaction costs, the contracts people write will be *incomplete* in important respects. The parties will quite rationally leave out many contingencies, taking the point of view that it is better to 'wait and see what happens' than to try to cover a large number of individually unlikely eventualities. Less rationally, the parties will leave out other contingencies that they simply do not anticipate. Instead of writing very long-term contracts the parties will write limited term contracts, with the intention of renegotiating these when they come to an end. (A paper which explores the implications of this is Crawford 1986). Contracts will often contain clauses which are vague or ambiguous, sometimes fatally so.

Anyone familiar with the legal literature on contracts will be aware that almost every contractual dispute that comes before the courts concerns a matter of incompleteness. In fact, incompleteness is probably at least as important empirically as asymmetric information as an explanation for departures from 'ideal' Arrow–Debreu contingent contracts. In spite of this, relatively little work has been done on this topic, the reason presumably being that an analysis of transaction costs is so complicated. One problem is that the first two transaction costs referred to above are intimately connected to the idea of bounded rationality (as in Simon 1982), a successful formalization of which does not yet exist. As a result, perhaps, the few attempts that have been made to analyse incompleteness have concentrated on the third cost, the cost of writing the contract.

One approach, due to Dye (1985), can be described as follows. Suppose that the amount of input, q, traded between a buyer and seller should be a function of the product price, p, faced by the buyer: $q = f(p)$. Writing down this function is likely to be costly. Dye measures the costs in terms of how many different values q takes on as p varies; in particular, if $\#\ \{q|q = f(p)$ for some $p\} = n$, the cost of the contract is $(n - 1)c$, where $c > 0$. This means that a noncontingent statement '$q = 5$ for all p' has zero cost, the statement '$q = 5$ for $p \leq 8$, $q = 10$ for $p > 8$' has cost c, and so on.

The costs Dye is trying to capture are real enough, but the measure used has some drawbacks. It implies for example, that the statement '$q = p^{1/2}$ for all p' has infinite cost if p has infinite domain, and does not distinguish between the cost of a simple function like this and the cost of a much more complicated function. As another example, a simple indexation clause to the effect that the real wage should be constant (i.e. the money wage $= \lambda p$ for some λ) would never be observed since, according to Dye's measure, it too has infinite cost. In addition, the approach does not tell us how to assess the cost of indirect ways of making q contingent; for example, the contract could specify that the buyer, having observed p, can choose any amount of input q he likes, subject to paying the seller σ for each unit.

There is another way of getting at the cost of including contingent statements. This is to suppose that what is costly is describing the state of the world ω rather than writing a statement per se. That is, suppose that ω cannot be represented simply by a product price, but is very complex and of high dimension – e.g., it includes the state of demand, what other firms in the industry are doing, the state of technology, etc. Many of these components may be quite nebulous. To describe the state *ex-ante* in sufficient detail that an outsider, e.g. The courts, can verify whether a particular state $\omega = \hat{\omega}$ has occurred, and so enforce the contract, may be prohibitively costly. Under these conditions, the contract will have to omit some (in extreme cases, all) references to the underlying state.

Similar to this is the case where what is costly is describing the *characteristics* of what is traded or the *actions* (e.g. investments) the parties must take. For example, suppose that there is only one

state of the world, but that q now represents the quality of the item traded rather than the quantity. An ideal contract would give a precise description of q. However, quality may be multidimensional and very difficult to describe unambiguously (and vague statements to the effect that quality should be 'good' may be almost meaningless). The result may be that the contract will have to be silent on many aspects of quality and/or actions.

Models of this sort of incompleteness have been investigated by Grossman and Hart (1987) and Hart and Moore (1985) for the case where the state of the world cannot be described and by Bull (1985) and Grossman and Hart (1986, 1987) for the case where quality and/or actions cannot be specified. These models do not rely on any asymmetry of information between the parties. Both parties may recognize that the state of the world is such that the buyer's benefit is high or the seller's cost is low, or that the quality of an item is good or bad or that an investment decision is appropriate or not. The difficulty is conveying this information to others. *That is, it is the asymmetry of information between the parties on the one hand, and outsiders, such as the courts, on the other, which is the root of the problem*.

To use the jargon, incompleteness arises because states of the world, quality and actions are *observable* (to the contractual parties) but not *verifiable* (to outsiders).

We describe an example of an incomplete contract along these lines in the next section.

Incomplete Contracts: An Example

We will give an example of an incomplete contract for the case where it is prohibitively costly to specify the quality characteristics of the item to be exchanged or the parties' investment decisions. Similar problems arise when the state of the world cannot be described. The example is a variant of the models in Grossman and Hart (1986, 1987), Hart and Moore (1985).

Consider a buyer B who wishes to purchase a unit of input from a seller S. B and S each make a (simultaneous) specific investment at date 0 and trade occurs at date 1. Let I_B, I_S denote, respectively, the investments of B and S, and to simplify assume that each can take on only two values, H or L (high or low). These investments are observable to B and S, but are not verifiable (they are complex and multidimensional, or represent effort decisions) and hence are non-contractible. We assume that at date 1 the seller can supply either 'satisfactory' input or 'unsatisfactory' input. 'Unsatisfactory' input has zero benefit for the buyer and zero cost for the seller (so it is like not supplying at all). 'Satisfactory' input yields benefits and costs which depend on *ex-ante* investments. These are indicated in Fig. 1.

The first component refers to the buyer's benefit, v, and the second to the seller's cost, c. So when $I_S = H$, $I_B = H$, $v = 10$ and $c = 6$ (if input is 'satisfactory'). From these gross benefits and costs must be subtracted investment costs, which we assume to be 1.9 if investment is high and zero if it is low (for each party). (All benefits and costs are in date 1 dollars.) Note that there is no uncertainty and so attitudes to risk are irrelevant.

Our assumption is that the characteristics of the input (e.g. whether it is 'satisfactory') are observable to both parties, but are too complicated to be specified in a contract. The fact that they are observable means that the buyer can be given the option to reject the input at date 1 if he does not like it. This will be important in what follows.

An important feature of the example is that the seller's investment affects not only the seller's costs but also the buyer's benefit and the buyer's investment affects not only the buyer's benefit but also the seller's costs. The idea here is that a better investment by the seller increases the quality of 'satisfactory' input; and a better investment by the buyer reduces the cost of producing 'satisfactory' input, that is input that can be used by the buyer.

For instance, one can imagine that B is an electricity generating plant and S a coal mine that the plant is sited next to. I_B might refer to the type of coal-burning boiler that the plant installs and I_S to the way the coal supplier develops the mine. By investing in a better boiler, the power plant may be able to burn lower quality coal, thus reducing the seller's costs, while still increasing its gross (of investment) profit. On the other hand, by developing a good seam, the coal

supplier may raise the quality of coal supplied while reducing its variable cost.

The first-best has $I_B = I_S = H$, with total surplus equal to $(10 - 6) - 3.8 = 0.2$ (if $I_B = H$ and $I_S = L$, or vice versa, surplus $= 0.1$ and if $I_B = I_S = L$, no trade occurs and surplus is zero). This could be achieved if *either* investment *or* quality were contractible as follows. If investment is contractible, an optimal contract would specify that the buyer must set $I_B = H$ and the seller $I_S = H$ and give the buyer the right to accept the input at date 1 at price p_1 or reject it at price p_0. If $10 > p_1 - p_0 > 6$, the seller will be induced to supply satisfactory input (the gain, $p_1 - p_0$, from having the input accepted exceeds the seller's supply cost) and the buyer to accept it (the buyer's benefit exceeds the increment price $p_1 - p_0$. If, on the other hand, quality is contractible, the contract could specify that the seller must supply input with the precise characteristics which make it satisfactory when $I_B = I_S = H$. Each party would then have the socially correct investment incentives since, with specific performance, neither party's investment affects the other's payoff (there is no externality).

We now show that the first-best cannot be achieved if investment and quality are both noncontractible. A second-best contract can make price a function of any variable that is verifiable. Investment and quality are not verifiable (nor is v or c), but we shall suppose that whether the item is accepted or rejected by the buyer is, so the contract can specify an acceptance price, p_1, and a rejection price, p_0. In fact, p_0, p_1 can also be made functions of (verifiable) messages that the buyer and seller send each other, reflecting the investment decisions that both have made (as in Hart and Moore 1985). The following argument is unaffected by such messages and so, for simplicity, we ignore them (the interested reader is referred to Hart and Holmstrom 1987).

Can we sustain the first-best by an appropriate choice of p_0, p_1? The seller always has the option of choosing $I_S = L$ and producing an item of unsatisfactory quality, which yields him a net payoff of p_0. In order to induce him not to do this, we must have

$$p_1 - 6 - 1.9 \geq p_0, \quad \text{i.e. } p_1 - p_0 \geq 7.9. \quad (2)$$

Similarly the buyer's net payoff must be no less than $-p_0$ since he always has the option of choosing $I_B = L$ and rejecting the input. That is,

$$10 - p_1 - 1.9 \geq -p_0, \quad \text{i.e. } p_1 - p_0 \leq 8.1. \quad (3)$$

So $(p - p_0)$ must lie between 7.9 and 8.1.

Now the seller has an additional option. If he expects the buyer to set $I_B = H$, he can choose $I_S = L$ and, given that $8.1 \geq p_1 - p_0 \geq 7.9$, still be confident that trade of 'satisfactory' input will occur under the original contract at date 1 (the buyer will accept satisfactory input since $v = 9 > p_1 - p_0$, while the seller will supply it since $p_1 - p_0 > 7 = c$). But if the seller deviates, his payoff rises from $p_1 - 6 - 1.9$ to $p_1 - 7$. (The example is symmetric and so a similar deviation is also profitable for the buyer.) Hence the $I_B = I_S = H$ equilibrium will be disrupted.

We see, then, that the first-best cannot be sustained if investment and quality are both noncontractible. The reason is that it will be in the interest of the seller (or the buyer) to reduce investment since, although this reduces social benefit by lowering the buyer's (or seller's) benefit, it increases the seller's (or buyer's) own profit. The optimal second-best contract will instead have $I_B = H$, $I_S = L$ (or vice versa), which will be sustained by a pair of prices p_0, p_1 such that $9 > p_1 - p_0 > 7$. Total surplus will be 0.1 instead of the first-best level of 0.2. (Note the importance of the assumption that both the buyer and seller can choose $I = H$ or L. If only the buyer (or the seller) can choose $I = L$, the first-best can be achieved by choosing $p_1 - p_0$ between 6 and 7 (or 9 and 10): any deviation by the buyer (or the seller) will then be unprofitable since it will lead to no trade.)

The conclusion is that inefficiencies can arise in incomplete contracts even though the parties have common information (both observe investments and both observe quality). The particular inefficiency that occurs in the model analysed is in *ex-ante* investments. *Ex-post* trade is always efficient relative to these investments since p_1, p_0 can and will be chosen such that $v > p_1 - p_0 > c$,

i.e. the seller wants to supply and the buyer to receive satisfactory input. The example can be regarded as formalizing the intuition of Williamson (1985) and Klein et al. (1978) that relationship-specific investments will be distorted due to the impossibility of writing complete contingent contracts – note that this result is achieved without imposing arbitrary restrictions on the form of the permissible contract (e.g. we have not ruled out the existence of long-term contracts from the start). (There is one exception to this statement – we have excluded the participation of a third party to the contract; for a discussion and justification of this, see Hart and Holmstrom 1987.)

The example may be used to illustrate a theory of ownership presented in Grossman and Hart (1986, 1987). It is sometimes suggested that when transaction costs prevent the writing of a complete contract, there may be a reason for firm integration (see Williamson 1985). Consider the payoffs of Fig. 1 and suppose that B takes over S. The control that B thereby gains over S's assets may allow B to affect S's costs in various ways, and this may reduce the possibility of opportunistic behaviour by S. To take a very simple (and contrived) example, suppose that if S chooses $I_S = L$, B can take some action, α with respect to S's assets at date 1 so as to make S's cost of supplying either satisfactory or unsatisfactory input equal to 9 (in the coal-electricity example, α might refer to the part of the mine's seam the coal is taken out of; note that we now drop the assumption that the cost of supplying unsatisfactory input is zero). Imagine furthermore that this action increases B's benefit, so that B will indeed take it at date 1 if S chooses L. Then with this extra degree of freedom, the first-best can be achieved. In particular, if $p_1 = p_0 + 6.1$, $I_S = I_L = H$ is a Nash equilibrium since, by the above reasoning, any deviation by the seller will be punished, while if the buyer deviates, the seller will supply unsatisfactory input given that $p_1 < p_0 + 7$.

	$I_B = H$	$I_B = L$
$I_s = H$	(10, 6)	(9, 7)
$I_s = L$	(9, 7)	(6, 10)

Incomplete Contracts, Fig. 1

Note that if action α could be specified in the initial contract, there would be no need for integration: the initial contract would simply say that B has the right to choose α at date 1. Ownership becomes important, however, if (i) α is too complicated to be specified in the date 0 contract and therefore qualifies as a residual right of control; and (ii) residual rights of control over an asset are in the hands of whomever owns that asset. The point is that under incompleteness the allocation of residual decision rights matters since the contract cannot specify precisely what each party's obligations are in every state of the world. To the extent that ownership of an asset guarantees residual rights of control over that asset, vertical and lateral integration can be seen as ways of ensuring particular – and presumably efficient – allocations of residual decision rights. (While in the above example, integration increases efficiency, this is in no way a general conclusion. In Grossman and Hart (1986, 1987), examples are presented where integration reduces efficiency.)

Before concluding this section, we should emphasize that for reasons of tractability we have confined our attention to incompleteness due to a very particular sort of transaction cost. In practice, some of the other transactions costs we have alluded to are likely to be at least as important, if not more so. For example, in the type of model we have analysed, although the parties cannot describe the state of the world or quality characteristics, they are still supposed to be able to write a contract which is unambiguous and which anticipates all eventualities. This is very unrealistic. In practice, a contract might, say, have B agreeing to rent S's concert hall for a particular price. But suppose S's hall then burns down. The contract will usually be silent about what is meant to happen under these conditions (there is no hall to rent, but should S pay B damages and if so how much?), and so, in the event of a dispute, the courts will have to fill in the 'missing provision'. (A situation where it becomes impossible or extremely costly to supply a contracted for good is known as one of

'impossibility' or 'frustration' in the legal literature.) An analysis of this sort of incompleteness, although extremely hard, is a very important topic for future research. It is likely to yield a much richer and more realistic view of the way contracts are written and throw light on how courts should assess damages (this latter issue has begun to be analysed in the law and economics literature; see, e.g., Shavell 1980).

Self-Enforcing Contracts

The previous discussion has been concerned with explicit binding contracts that are enforced by outsiders, such as the courts. Even the most casual empiricism tells us that many agreements are not of this type. Although the courts may be there as a last resort (the shadow of the law may therefore be important), these agreements are enforced on a day to day basis by custom, good faith, reputation, etc. Even in the case of a serious dispute, the parties may take great pains to resolve matters themselves rather than go to court. This leads to the notion of a self-enforcing or implicit contract (the importance of informal arrangements like this in business has been stressed by Macaulay (1963) and Ben-Porath (1980) among others).

People often by-pass the legal process presumably because of the transaction costs of using it. The costs of writing a 'good' long-term contract discussed in section "The Costs of Writing Long-Term Contracts" are relevant here. So also is the skill with which the courts resolve contractual disputes. If contracts are incomplete and contain missing provisions as well as vague and ambiguous statements, appropriate enforcement may require abilities and knowledge (what was in the parties' minds?) that many judges and juries do not possess. This means that going to court may be a considerable gamble – and an expensive one at that. (This is an example of the fourth transaction cost noted in section "The Costs of Writing Long-Term Contracts".)

Although the notion of implicit or self-enforcing contracts is often invoked, a formal study of such agreements has begun only recently (see, e.g. Bull 1985), with a considerable stimulus coming from the theory of repeated games. This literature has stressed the role of *reputation* in 'completing' a contract. That is, the idea is that a party may behave 'reasonably' even if he is not obliged to do so in order to develop a reputation as a decent and reliable trader. In some instances such reputational effects will operate only within the group of contractual parties – this is sometimes called *internal* enforcement of the contract – while in others the effects will be more pervasive. The latter will be the case when some outsiders to the contract, for example other firms in the industry or potential workers for a firm, observe unreasonable behaviour by one party, and as a result are more reluctant to deal with it in the future. In this case the enforcement is said to be *external* or *market-based*. Note that there may be a tension between this external enforcement and the reasons for the absence of a legally binding contract in the first place – the more people can observe the behaviour, the more likely it is to be verifiable.

The distinction between an incomplete contract and a standard asymmetric information contract should be emphasized here. It is the former that allows reputation to operate since the parties have the same information and can observe whether reasonable behaviour is being maintained. In the latter case, it is unclear how reputation can overcome the asymmetry of information between the parties that is the reason for the departure from an Arrow–Debreu contract.

The role of reputation in sustaining a contract can be illustrated using the following model (based on Bull (1985) and Kreps (1984); this is an even simpler model of incomplete contracts than that of the last section). Assume that a buyer, B, and a seller, S, wish to trade an item at date 1 which has value v to the buyer and cost c to the seller, where $v > c$. There are no *ex-ante* investments and the good is homogeneous, so quality is not an issue. Suppose, however, that it is not verifiable whether trade actually occurs. Then a legally binding contract which specifies that the seller must deliver the item and the buyer must pay p, where $v > p > c$, cannot be enforced. The reason is that, assuming (as we shall) that simultaneous delivery and payment are infeasible,

if the seller has to deliver first, the buyer can always deny that delivery occurred and refuse payment, while if the buyer has to pay first, the seller can always claim later that he did deliver even though he did not. As a result, if the parties must rely on the courts, a gainful trading opportunity will be missed.

The idea that not even the level of trade is verifiable is extreme, and Bull (1985) in fact makes the more defensible assumption that it is the quality of the good that cannot be verified (in Bull's model, S is a worker and quality refers to his performance). Bull supposes that quality is observable to the buyer only with a lag, so that take it or leave it offers of the type considered in the last section are not feasible. As a result the seller always has an incentive to produce minimum quality (which corresponds in the above model to zero output). Making quantity nonverifiable is a cruder but simpler way of capturing the same idea (this is the approach taken in Kreps 1984).

Note that in the above model incompleteness of the contract arises entirely from transaction cost (3), the difficulty of writing and enforcing the contract.

To introduce reputational effects one supposes that this trading relationship is repeated. Bull (1985) and Kreps (1984) follow the supergame literature and assume infinite repetition in order to avoid unravelling problems. This approach, as is well known, suffers from a number of difficulties. First, the assumption of infinite (or in some versions, *potentially* infinite) life is hard to swallow. Secondly, 'reasonable' behaviour, i.e. trade, is sustained by the threat that if one party behaves unreasonably so will the other party from then on. While this threat is 'credible' (more precisely, subgame perfect), it is unclear why the parties could not decide to continue to trade after a deviation, i.e. to 'let bygones be bygones' (see Farrell 1984.)

It would seem that a preferable approach is to assume that the relationship has finite length, but introduce asymmetric information, as in Kreps and Wilson (1982) and Milgrom and Roberts (1982). The following is based on some very preliminary work that Bengt Holmstrom and I have undertaken along these lines.

Suppose that there are two types of buyers in the population, honest and dishonest. Honest buyers will always honour any agreement or promise that they have made while dishonest ones will do so only if this is profitable. A buyer knows his own type, but others do not. It is common knowledge that the fraction of honest buyers in the population is π, $0 < \pi < 1$. In contrast, all sellers are known to be dishonest. All agents are risk neutral.

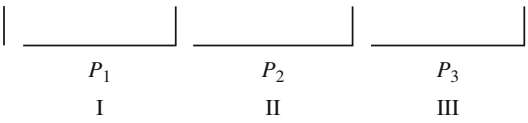
Assume for simplicity that a single buyer and seller are matched at date 0 with neither having any alternative trading partners at this date or in the future. Consider first the one-period case. Then a date 0 agreement can be represented as follows. The interpretation is that the buyer promises to pay the seller p_1 before date 1 (stage I); in return, the seller promises to supply the item at date 1 (stage II); and in return for this, the buyer promises to make a further payment of p_2 (stage III).

We should mention one further assumption. Honest buyers, although they never breach an agreement first, are supposed to feel under no obligation to fulfil the terms of an agreement that has already been broken by a seller (interestingly, although this is a theory of buyer psychology, it has parallels in the common law). Note that if a buyer ever breaks an agreement first, he reveals himself to be dishonest, with the consequence that no further self-enforcing agreement with the seller is possible and hence trade ceases.

What is an optimal agreement? Consider Fig. 2. The seller knows that he will receive p_2 only with probability π since a dishonest buyer will default at the last stage. Since the seller is himself dishonest, he will supply at Stage II only if it is profitable for him to do so, i.e. only if

$$\pi p_2 - c \geq 0. \tag{4}$$

Assume for simplicity that the seller has all the bargaining power at date 0 (nothing that follows



Incomplete Contracts, Fig. 2

depends on this). Then the seller will wish to maximize his overall payoff

$$p_1 + \pi p_2 - c, \tag{5}$$

subject to (4) which makes it credible that he will supply at stage II and also the constraint that he does not discourage an honest buyer from participating in the agreement at date 0. Since with (4) satisfied, buyers know that they will receive the item for sure, this last condition is

$$v - p_1 - p_2 \geq \quad 0. \tag{6}$$

Note that a dishonest buyer's payoff $v - p_1$ is always higher than an honest buyer's payoff given in (6), so there is no way to screen out dishonest buyers. In the language of asymmetric information models, the equilibrium is a pooling one.

Since the seller's payoff is increasing in p_1, (6) will hold with equality (the buyer gets no surplus). (More generally, changes in p_1 simply redistribute surplus between the two parties without changing either's incentive to breach.) If we substitute for p_1 in (5), the seller's payoff becomes $v - p_2(1 - \pi) - c$, which, when maximized subject to (4), yields the solution $p_2 = c/\pi$. The maximized net payoff is

$$v - c/\pi, \tag{7}$$

which is less than the first-best level, $v - c$.

We see then that the conditions for trade are more stringent in the absence of a binding contract. If $c/(\pi) > v > c$, there are gains from trade which would not be realized in a one-period relationship.

Suppose now that the relationship is repeated. Consider a two-period version of the above and assume no discounting. Now the diagram shown in Fig. 3 applies. That is, the agreement says that the buyer pays, the seller supplies the first time, the buyer pays more, the seller supplies a second time, and the buyer makes a final payment. Rather than solving for the optimal arrangement, we shall simply show that the seller can do better than in the one period case. Let $p_3 = c/\pi$, $p_2 = c$ and $p_1 = 2v - c - (c)/\pi$. Then (i) the seller will supply at Stage IV (if matters have got that far), knowing that he will receive $p3$ with probability π (ii) both honest and dishonest buyers will pay p_2 at Stage III, the latter because, at a cost of c, they thereby ensure supply worth $v > c$ at Stage IV; (iii) the seller will supply at stage II because this gives him a net payoff of $p_2 + \pi p_3 - 2c \geq 0$, while if he does not the arrangement is over and his payoff is zero; (iv) an honest buyer is prepared to participate since his surplus is non-negative (actually zero).

The seller's overall expected net payoff is

$$p_1 + \pi_2 + \pi p_3 - 2c = 2v - c - c/\pi, \tag{8}$$

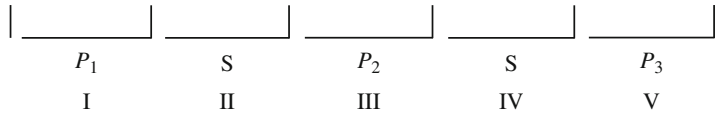
which *exceeds* twice the one-period payoff. Hence trade is more likely to take place in a two-period relationship than in a one-period one. In fact it can be shown that the above is an *optimal* two-period agreement.

Repetition improves things by allowing the honest buyer to pay less second time round (Stage III) than third time round (Stage V). That is, the arrangement *back-loads* payments. This is acceptable to the seller because he knows that even a dishonest buyer will not default at Stage III since he has a large stake in the arrangement continuing. To put it another way, the dishonest buyer does not want to reveal his dishonesty at too early a stage.

The same arrangement can be used when there are more than two periods: the buyer promises to pay c at every stage except the last, when he pays (c/π). In fact the per period surplus of the seller from such an arrangement converges to the first-best level $(v - c)$ as the number of periods tends to ∞ (assuming no discounting, of course).

Although the above analysis is extremely provisional and sketchy, we can draw some tentative

Incomplete Contracts, Fig. 3



conclusions about the role of reputation and indicate some directions for further research. First, the notion of a psychic cost of breaking an agreement seems to be a useful – as well as a not unrealistic – basis for a theory of self- enforcing contracts. It is obviously desirable to drop the assumption that some agents are completely honest and others completely dishonest, and assume instead that the typical trader has a finite psychic cost of breaking an agreement, where this cost is distributed in the population in a known way. In other words, everybody 'has their price', but this price varies. Preliminary work along these lines suggests that the above results generalize; in particular, repetition makes it easier to sustain a self-enforcing agreement.

Of course, asymmetries of information about psychic costs are not the only possible basis for a theory of reputation. For example, the buyer and seller could have private information about v and c, and might choose their trading strategies to influence perceptions about the values of these variables (the role of uncertainty about v and c in determining reputation has been investigated by Thomas and Worrall 1984). A theory of self-enforcing contracts should ideally generate results which are not that sensitive to where the asymmetry of information is placed. The work of Fudenberg and Maskin (1986) in a related context, however, suggests that this may be a difficult goal to achieve.

There are a number of other natural directions in which to take the model. One is to introduce trade with other parties. For example, the seller may trade with a succession of buyers rather than a single one. The extent to which repetition increases per period surplus in this case depends on whether new buyers observe the past broken promises of the seller. (This determines the degree to which external enforcement operates; more generally, 'a new buyer may observe that default occurred in the past, but be unsure about who was responsible for it.) If new buyers do not observe past broken promises, repetition achieves nothing, which gives a very strong prediction of the possible benefits of a long-term relationship between a fixed buyer and seller. Even if past broken promises are observed perfectly, it appears that, *ceteris paribus*, a single long-term agreement may be superior to a succession of short-term ones. The reason is that in the latter case the constraint is imposed that each party must receive non-negative surplus over *their* term of the relationship whereas in the former case there is only the single constraint that surplus must be non-negative over the whole term (see Bull 1985; Kreps 1984).

Probably the most important extension is to introduce incompleteness due to other sorts of transaction costs, e.g. the 'bounded rationality' costs (1) and (2) discussed in section "The Costs of Writing Long-Term Contracts". The problem is that the same factors which make it difficult to anticipate and plan for eventualities in a formal contract apply also to informal arrangements. That is, an informal arrangement is also likely to contain many 'missing provisions'. But then the question arises, what constitutes 'reasonable' or 'desirable' behaviour (in terms of building a reputation) with regard to states or actions that were not discussed *ex-ante*? Custom, among other things, is likely to be important under these conditions: behaviour will be 'reasonable' or 'desirable' to the extent that it is generally regarded as such (for a good discussion of this, see Kreps 1984). This raises many new and interesting (as well as extremely difficult) questions.

Summary and Conclusions

The vast majority of the theoretical work on contracts to date has been concerned with what might be called 'complete' contracts. In this context, a complete contract means one that specifies each party's obligations in every conceivable eventuality, rather than a contract that is fully contingent in the Arrow–Debreu sense. In particular, according to this terminology, the typical asymmetric information contract found in the principal-agent or implicit contract literatures (see Hart and Holmstrom 1987) is complete.

In reality it is usually impossible to lay down each party's obligations completely and unambiguously in advance, and so most actual contracts are seriously incomplete. In this entry, we have

tried to indicate some of the implications of such incompleteness. Among other things, we have seen that incompleteness can lead to departures from the first-best even when there are no asymmetries of information among the contracting parties (and, moreover, the parties are risk neutral).

More important perhaps than this is the fact that incompleteness raises new and difficult questions about how the behaviour of the contracting parties is determined. To the extent that incomplete contracts do not specify the parties' actions fully, i.e. they contain 'gaps', additional theories are required to tell us how these gaps are filled in. Among other things, outside influences such as custom or reputation may become important under these conditions. In addition, outsiders, such as the courts (or arbitrators), may have a role to play in filling in missing provisions of the contract and resolving ambiguities rather than in simply enforcing an existing agreement. Incompleteness can also throw light on the importance of the allocation of decision rights or rights of control. If it is too costly to state precisely how a particular asset is to be used in every state of the world, it may be efficient simply to give one party 'control' of the asset, in the sense that he is entitled to do what he likes with it, subject perhaps to some explicit (contractible) limitations.

While the importance of incompleteness is very well recognized by lawyers, as well as by those working in law and economics, it is only beginning to be appreciated by economic theorists. It is to be hoped that work in the next few years will lead to significant advances in our formal understanding of this phenomenon. Unfortunately, progress is unlikely to be easy since many aspects of incompleteness are intimately connected to the notion of bounded rationality, a satisfactory formalization of which does not yet exist.

As a final illustration of the importance of incompleteness, consider the following question. Why do parties frequently write a limited term contract, with the intention of renegotiating this when it comes to an end, rather than writing a single contract that extends over the whole length of their relationship? In a complete contract framework such behaviour cannot be advantageous since the parties could just as well calculate what will happen when the contract expires and include this as part of the original contract. It is to be hoped that future work on incomplete contracts will allow this very basic question to be answered.

See Also

- ▶ Adverse Selection
- ▶ Contract Theory
- ▶ Exchange
- ▶ Implicit Contracts
- ▶ Moral Hazard
- ▶ Rationality, Bounded

Bibliography

Becker, G. 1964. *Human capital*. New York: Columbia University Press.

Ben-Porath, Y. 1980. The F-connection: Families, friends, and firms and the organization of exchange. *Population and Development Review* 6: 1–30.

Bull, C. 1985. *The existence of self-enforcing implicit contracts*. New York: C.V. Starr Center, New York University.

Crawford, V. 1986. *Long-term relationships governed by short-term contracts*. Princeton: Princeton University.

Dye, R. 1985. Costly contract contingencies. *International Economic Review* 26 (1): 233–250.

Freixas, X., R. Guesnerie, and J. Tirole. 1985. Planning under incomplete information and the ratchet effect. *Review of Economic Studies* 52(2): 169, 173–192.

Fudenberg, D., and E. Maskin. 1986. The Folk Theorem in repeated games with discounting and with incomplete information. *Econometrica* 54 (3): 533–554.

Goldberg, V., and J. Erickson. 1982. *Long-term contracts for petroleum coke*. Department of Economics Working Paper Series No. 206, University of California, Davis, September.

Grossman, S., and O. Hart. 1986. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* 94 (4): 691–719.

Grossman, S., and O. Hart. 1987. Vertical integration and the distribution of property rights. In *Economic policy in theory and practice*, Sapir conference volume. London: Macmillan Press.

Grout, P. 1984. Investment and wages in the absence of binding contracts: A Nash bargaining approach. *Econometrica* 52 (2): 449–460.

Hart, O., and B. Holmstrom. 1987. The theory of contracts. In *Advances in economic theory, fifth World Congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

Hart, O., and J. Moore. 1985. *Incomplete contracts and renegotiation*. London School of Economics, Working Paper.
Joskow, P. 1985. Vertical integration and long-term contracts. *Journal of Law, Economics and Organization* 1, Spring.
Klein, B., R. Crawford, and A. Alchian. 1978. Vertical integration, appropriable rents and the competitive contracting process. *Journal of Law and Economics* 21: 297–326.
Kreps, D. 1984. *Corporate culture and economic theory.* Mimeo: Stanford University.
Kreps, D., and R. Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27: 253–279.
Kydland, F., and E. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85 (3): 473–492.
Macaulay, S. 1963. Non-contractual relations in business: A preliminary study. *American Sociological Review* 28: 55–67.
Milgrom, P., and D.J. Roberts. 1982. Predation, reputation and entry deterrence. *Journal of Economic Theory* 27: 280–312.
Shavell, S. 1980. Damage measures for breach of contract. *Bell Journal of Economics* 11 (2): 466–490.
Simon, H. 1982. *Models of bounded rationality.* Cambridge, MA: MIT Press.
Thomas, J., and T. Worrall. 1984. *Self-enforcing wage contracts*. Mimeo: University of Cambridge.
Williamson, O. 1985. *The economic institutions of capitalism*. New York: Free Press.

Incomplete Markets

Charles Wilson

Abstract

'Incomplete markets' describes a market structure in which there are effective constraints on which bundles of goods may be exchanged with each other. When incompleteness arises from markets that are sequentially segmented, some of the basic properties of general equilibrium are affected. First, equilibrium may not exist even under the usual regularity assumptions. Second, allocations may not be Pareto optimal, even after the limitations imposed by the market structure are taken into account. Third, if securities are denominated in nominal values, the equilibrium allocation is generally not locally unique.

Keywords

Arrow–Debreu markets; Complete markets; Demand functions; Existence of equilibrium; Incomplete markets; Indeterminacy of equilibrium; Local uniqueness of equilibrium; Marginal rate of substitution; Spot markets; Uncertainty

JEL Classifications

D4

Incomplete markets arise when agents are unable to exchange every good either directly or indirectly with every other agent. In the case of a single market with no limitations on the exchange of goods, relatively mild assumptions guarantee the existence, Pareto optimality, and local uniqueness of a competitive equilibrium. However, once we impose restrictions on the trade of goods and introduce sequential markets so that not all trade take place in a single market, any one of these properties may fail to be satisfied. A large literature has evolved that examines the conditions under which different sets of securities generate a complete set of markets and the properties of the equilibrium allocations when they do not. In this article I illustrate a few of the main ideas in this literature.

A good starting point is the work of Arrow (1973), who demonstrated that static competitive analysis can be extended to deal with the case of uncertainty, but only by expanding the set of markets to include a separate price for each commodity in each state of the world. With a complete set of markets in state contingent commodities, it follows immediately that, under the usual conditions on preferences, the competitive allocation exists, is *ex-ante* Pareto efficient and locally unique. Although this approach solves the problem of extending general equilibrium analysis to deal with uncertainty, it strains the credibility of the model by requiring an unrealistic number of goods to be simultaneously exchanged. It is important, therefore, to examine the extent to

which the same allocation can be attained with a different market structure requiring a smaller number of instruments.

Consider an economy with S possible realizations of an uncertain state of the world with N goods in each state. We will refer to the state contingent good i in state s as good is. To allow for more general market structures, we suppose that all trading takes place in securities, which are claims on the vector of state contingent goods. A *simple* security promises delivery of one unit of only one state contingent good. Observe that any security may be represented as a linear combination of simple securities. The *span* of any set of securities is the set of state-contingent goods that can be obtained by some linear combination of those securities. A *market* is a set of securities and a price vector at which they may be exchanged. An Arrow–Debreu market is a market consisting of the complete set of simple securities, and an Arrow–Debreu allocation is the competitive equilibrium for an Arrow–Debreu market. A *spot* market for state s is a market in which only the simple securities for s goods are traded. We will always assume that the spot market for each state s is 'complete' in the sense that the set of feasible trades spans the set of all simple securities for state s.

To reduce the number of securities traded in any market, Arrow considers a two-stage market structure. In the first stage, before the state is realized, all agents have access to a 'securities' market. In this market, there is one security f^s for each state s, which represents a claim of one unit of each good in that state. In the second stage, after the state is realized and the claims of the first-stage securities are realized, the corresponding spot market opens and the final allocation is determined by the spot market equilibrium. Arrow demonstrates that, when agents have perfect foresight of the future spot market prices, any Arrow–Debreu allocation can be attained as a competitive equilibrium for this two-stage market structure. Since spot markets operate only when the actual state is realized, the total number of securities that are required to obtain the Arrow–Debreu allocation is reduced from NS to N + S.

To demonstrate the logic of Arrow's result, let p(s) denote the vector of spot prices in state s and let qs denote the first-stage price of security fs. Then, defining pis = qspi(s) for each good is, we obtain an NS vector (p_{is}) that defines the relative prices for all state contingent goods. For example, to exchange good is for good js', an agent exchanges security $f^{s'}$ for security f^s in the first-stage securities market and then uses the spot markets to obtain the desired net exchange. Alternatively, given a vector (p_{is}) of state-contingent prices, we may obtain the equivalent prices for the two-stage market by defining each $q_s = \Sigma_i p_{is}$ and each $p_i(s) = p_{is}/q_s$. Then, since each agent effectively faces the same budget constraint in both market structures, it follows that both market structures generate the same equilibrium allocation of goods.

Notice that the only role of the first-stage securities market is to transfer purchasing power across states. For instance, the set of simple securities of good 1 would work just as well. The essential requirement is that the set of securities spans the set of all possible transfers of purchasing power across states. Furthermore, so long as there is an 'insurance' security for each state s that delivers only state s contingent goods, the spanning condition is necessarily satisfied (at least if the vectors of all spot prices are strictly positive). Other sets of securities may also satisfy the spanning condition. However, the consideration of a more general set of securities also introduces some complications that may impact on the existence of equilibrium and the welfare analysis of the market structure.

The problem is not just that the set of available securities might not span the entire commodity space. In fact, as long as all trade takes place in a single market so that any feasible security can be traded with another, the span of the market is fixed. So if we simply redefine the commodity space as the market span and restrict preferences accordingly, the existence, Pareto optimality, and local uniqueness of equilibria (for any basis of securities) follow from the standard arguments. With multi-stage markets, however, the security markets are essentially segmented. Consequently, a change in relative spot market prices may translate into changes in the space of feasible transfers

of purchasing power that may be obtained by exchanging any given set of securities.

To illustrate, suppose there are only two goods in each state and the first-stage securities market consists of just two 'forward' securities, which respectively represent the claim of one unit of good X or one unit of good Y regardless of the state. Now fix the spot market prices in each state. Then since there are only two securities, it follows immediately that the dimension of the space of income transfers (measured, say, in terms of good X in each state) that can be obtained using the first- stage securities is at most two. Furthermore, if there are more than two states, the space of transfers that are spanned by the securities market depends on the relative prices in the different spot markets. For instance, suppose that the relative spot prices are the identical in states 1 and 2. Then any transfer of income to state 1 must be accompanied by the same transfer of income to state 2. However, if $p_x(1)/p_y(1) < 1 < p_x(2)/p_y(2)$, income can be transferred from state 1 to state 2 by exchanging one of forward security X for one unit of forward security Y. For the general case with N goods and S states, Townsend (1978) shows that when all first- stage securities are forward securities, the income transfers of these securities span R^S, the space of income transfers, if and only if there are at least S securities and the set of spot market price vectors are linearly independent.

The Existence of Equilibrium

When the dimension of the span of the transfers of a set of securities depend on the prices in the spot markets, the usual regularity assumptions on preferences no longer guarantee the existence of an equilibrium. Consider the following example based on Hart (1975). There are two agents, a and b, and two states. In each state there are two goods, labelled X and Y which must be consumed in non-negative amounts by each agent. The preferences and endowments of the agents are given in Table 1, where x_{as} and y_{as} are the respective amounts of good X and Y consumed by agent α in state s.

Incomplete Markets, Table 1

Agent	Endowments		Utility
	(X_1, Y_1)	(X_2, Y_2)	
a	(2, 2)	(1, 1)	$3x_{a1} + y_{a1} + 3x_{a2} + y_{a2}$
b	(1, 1)	(2, 2)	$x_{b1} + 3y_{b1} + x_{b2} + 3y_{b2}$

Agent a is endowed with two units of each good in state 1 and one unit of each good in state 2. His marginal rate of substitution between X and Y in either state is 3, and his marginal rate of substitution between goods across states is 1. Agent b is endowed with one unit of each good in state 1 and two units of each good in state 2. His preferences are the same as those of agent a except that the role of X and Y is reversed. It is easy to check that in the unique Arrow–Debreu equilibrium, the price of all state contingent goods must be equal.

Consider next the case in which the first-stage market consists of the two forward securities. We will show that a competitive equilibrium does not exist. As above, let $p_x(s)$ and $p_y(s)$ denote the equilibrium prices of goods X and Y in the state s spot market, and let q_x and q_y denote the equilibrium prices of the two forward securities. Now suppose some agent α exchanges q_x units of security Y for q_y units of security X. Then his income in state s changes by the amount $p_{xs}q_y - p_{ys}q_x$. Therefore, in equilibrium either (a) q_x/q_y lies between $p_x(1)/p_y(1)$ and px(2)/py(2), or (b) qx/qy = px1/py1 = px2/py2. Otherwise, one security dominates the other in the sense that an exchange of securities raises or lowers purchasing power in both states.

We show first that case (b) in which the relative spot prices are equal is not consistent with equilibrium. In this case, an exchange of securities leaves the income in both spot markets unchanged. Consequently, the equilibrium allocation and prices in the spot market must be the same as if no securities market existed. But the solution to either spot market then yields the allocation in which agent a obtains all three units of good X and agent b all three units of good Y. However, to clear the spot markets, the spot prices in the two states must differ, with $p_{x1}/p_{y1} = 2$ and

$p_{x2}/p_{y2} = 1/2$. We conclude that the relative spot prices cannot be equal in equilibrium.

Now suppose the relative spot prices are different. Then, using both the securities market and the spot markets, an agent may exchange good X in state 1 for good X in state 2 at the relative price $(p_x(1)/p_y(1))([p_x(2)q_y - p_y(2)q_x]/[p_y(1)q_x - p_x(1)q_y])$. Since markets are now effectively complete, the equilibrium prices in the market structure with forward securities must generate an Arrow–Debreu allocation. But we have already observed that the prices of state contingent goods must all be equal in an Arrow–Debreu equilibrium. It then follows that the relative spot prices in the two states must also be equal, which contradicts our conclusion above. We conclude that there is no competitive equilibrium for the forward security market structure.

In this example, an equilibrium fails to exist for the market structure with forward securities because the dimension of the resulting space of feasible net trades in state contingent goods abruptly shrinks at certain prices. As the relative prices of future securities and spot prices converge to the same ratio, the volume of trade in future securities that is required for a given transfer of purchasing power across states goes to infinity. Consequently, the demand functions for securities may be unbounded even in regions where all relative prices are bounded away from zero. To avoid this problem, Radner (1972) imposes an exogenous lower bound on short sales of securities and shows that this is sufficient to guarantee the existence of equilibrium under standard assumptions. Another approach is to assume that the set of securities is sufficiently rich to guarantee that the dimensionality of net trades does not vary with the price as in Geanokoplos and Polemarchakis (1986). Under these conditions, the demand functions remain bounded and continuous, so there is no need for an exogenous lower bound on excess demand. Kreps (1979) also notes that the set of transfers for any set of securities has full rank for almost all spot prices and therefore that the existence problem is not generic. A general theorem for the generic existence of equilibrium is established by Duffie and Shafer (1985).

Pareto Efficiency

As observed above, whenever the two-stage market structure generates a complete set of markets, the equilibrium allocation is an Arrow–Debreu allocation and is therefore Pareto optimal. However, if the first-stage market does not span the space of income transfers, then markets are not complete and the equilibrium allocation is generally not Pareto optimal. In this case, it may be of more use to restrict attention to a more limited set of allocations that reflect the restrictions imposed by the market structure. With the segmentation of markets, however, it is not immediately obvious how we should redefine the set of feasible allocations. For instance, if we permit a central planner to reallocate securities in each spot market, then any technologically feasible allocation can be obtained. To capture the restrictions implied by the market structure, therefore, we must impose some restrictions on how the spot market securities may be allocated.

One possibility is to permit the central planner to arbitrarily allocate securities in the first-stage market, but leave the allocation of securities in the spot markets to be determined by market clearing prices. This approach leads to the following definition suggested by Hart (1975). Let F denote the set of securities in the first- stage market. An allocation of state contingent goods is *constrained Pareto efficient* if (*a*) it is attained as an equilibrium in the spot markets for some feasible distribution of securities in F, and (*b*) there is no Pareto superior allocation of state contingent goods attained as an equilibrium in the spot markets for some other feasible distribution of securities in F.

We will show that when the number of securities in F is less than S, an equilibrium need not be even constrained Pareto efficient. The reason is that a redistribution of the ownership of securities generally leads to a change in the spot market prices and hence to a change in the vector of income transfers associated with each security. As we observed above, when the set of securities in F does not span R^S, the space of transfer vectors that are spanned by the securities in F generally depends on the spot market prices. Consequently, the transfer of real income generated by the

Incomplete Markets, Table 2 Caption missing

Agent	Endowments		Utility
	(X_1, Y_1)	(X_2, Y_2)	
a	(0, 2)	(2, 0)	$x_{a1} + \varepsilon \min \{x_{a2}, y_{a2}\}$
b	(2, 0)	(0, 2)	$\varepsilon \min \{x_{b1}, y_{b1}\} + x_{b2}$
c	(1, 1)	(1, 1)	$y_{c1} + y_{c2}$

redistribution of securities following the adjustment of prices in the spot markets typically lies outside the span of the transfers generated by the set of securities at the competitive equilibrium prices. By redistributing existing securities, therefore, it may be possible to increase the welfare of every agent in the economy.

To illustrate, consider an economy with three agents, *a, b* and *c*, and two states of the world, 1 and 2. In each state *s* there are two goods, labelled *X* and *Y*. Suppose the preferences and endowments of the agents are given by Table 2, where $x_{\alpha s}$ and $y_{\alpha s}$ are the respective consumption of goods *X* and *Y* by agent α in state *s*.

In this economy, agent *a* is endowed with two units of good *Y* in state 1 and two units of good *X* in state 2. He consumes only good *X* in state 1 and always consumes an equal amount of both goods in state 2. For each pair of units of the two goods he consumes in state 2 he is willing to give up ε units of his consumption of good *X* in state 1. The endowment and preferences of agent *b* are the same except that the role of the two states is reversed. Agent *c* is endowed with one unit of good *X* in both states but consumes only good *Y*. His marginal rate of substitution between consumption in the two states is one.

Suppose there is a single security that promises to deliver one unit of good *X* in each state. Since there is nothing for which to exchange this security, the equilibrium income and spot prices in each state will be determined solely by the endowments of the agents in that state. It is easy to check that the relative price of the two goods is one in both states. Agent *a* consumes two units of good *X* in state 1 and one unit of each good in state 2. Agent *b* consumes one unit of each good in state 1 and one unit of good *X* in state 2. Agent *c* consumes two units of good *Y* in both states.

Incomplete Markets, Table 3 Caption missing

Agent	Endowments	
	(X_1, Y_1)	(X_2, Y_2)
a	(−2, 2)	(0, 0)
b	(0, 0)	(−2, 2)
c	(5, 1)	(5, 1)

Although the security will never be traded in the market, it can still be used by the government to redistribute purchasing power in the two states and thereby change the spot prices. Suppose, for instance, that agents *a* and *b* must each supply agent *c* with two units of the security. Then the effect is the same as if the endowments were changed as to the endowments listed in Table 3.

For this economy the equilibrium price of good *Y* in terms of good *X* in each state is 5/2. Agent *a* consumes the three units of good *X* in state 1 and nothing in state 2. Agent *b* consumes nothing in state 1 and all three units of good *X* in state 2. Agent *c* consumes the three units of good *Y* in both states.

Now compare the welfare of the two agents in the two economies. Without the transfer payments, agents *a* and *b* attain an expected utility of $2 + \varepsilon$ while agent c attains an expected utility of 4. With the transfer payments, agent *a* and *b* both attain an expected utility of 3 while agent c attains a utility of 6. Consequently, for $0 < \varepsilon < 1$, the equilibrium with transfer payments Pareto dominates the equilibrium without transfer payments. By transferring purchasing power to agent *c* in both states, the economy has made the price of the goods demanded by agents *a* and *b* cheaper in those states where they value their increased welfare the most.

The possibility that securities can be reallocated to attain a Pareto superior allocation when markets are incomplete was first illustrated

by Hart (1975). He provided an example in which removing securities and hence decreasing the possibilities for trade actually resulted in a Pareto superior allocation. The intuition is similar to that provided in the example above. If markets are not compete, the introduction of a new security may change the spot market prices in such a way that utilities of all agents decrease unless they can make trades that are not available with the existing set of securities.

Geanokoplos and Polemarchakis (1986) consider a model with two periods and enlarge the commodity space to include consumption before the state of nature is realized. With a complete set of spot markets in the second period and a combined spot and securities market in the first period, they establish that the competitive equilibrium is almost never constrained Pareto optimal whenever the number of securities in F is less than S and there are at least two goods in each state. Geanokoplos et al. (1990) establish a similar result for a general equilibrium model of the stock market.

Nominal Securities and the Indeterminacy of Equilibrium

Cass (1985) investigates the implications for equilibrium when some of the securities are 'nominal'. These are securities in which the returns in any state are denominated in some unit of account. When all securities are nominal an equilibrium always exists. However, if the dimension of the span of these securities is less than S, the equilibrium is generally not locally unique. In fact, the dimension of indeterminacy is generally equal to $S - 1$.

This result derives from the fact that the real income actually transferred to any state by a nominal security depends on the price level in that state. Suppose the prices in each spot market s are normalized so that they sum to q_s. Then for each vector, $q = (q_1, \ldots q_S)$, any given nominal security f that promises delivery of f_s units of income in each state s corresponds to a unique 'real' security which pays f_s/q_s units of each good in each state s. Let $g_{f(q)}$ denote the real security $(f_1/q_1, \ldots, f_S/q_S)$, and let $F(q)$ denote the set of all such securities generated by the initial set of nominal securities. Then, for any security in $F(q)$, the relative prices in each state do not affect the amount of real income that is transferred by any given exchange of securities. Consequently, there will generally be a locally unique equilibrium associated with each vector q.

Suppose that the set of nominal securities does not span R^S. Then, any non-proportional change in q (generically) changes the span of $F(q)$. Consequently, when we replace the market of nominal securities with a market of real securities $F(q)$, each (normalized) vector q generally produces a distinct equilibrium allocation. Observe, however, that each of these allocations can be realized as an equilibrium with the same set of nominal securities. Therefore, since the dimension of normalized vectors q is $S - 1$, it follows that the dimension of equilibrium allocations associated with any incomplete set of nominal securities is generically $S - 1$.

Notice that this argument only works when the set of nominal securities does not span RS. When the span is complete, the possibilities for distributing real income using the artificial real securities no longer depend on q. Consequently, any equilibrium must yield an Arrow–Debreu allocation.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Multiple Equilibria in Macroeconomics
- ▶ Uncertainty and General Equilibrium

Bibliography

Arrow, K. 1973. The role of securities in the optimal allocation of risk-bearing. In *Essays in the theory of risk-bearing*. Chicago: Markham.

Cass, D. 1985. On the 'number' of equilibrium allocations with incomplete financial markets. Working Paper No. 85–16, CARESS, University of Pennsylvania.

Duffie, D., and W. Shafer. 1985. Equilibrium with incomplete markets, I: A basic model of generic existence. *Journal of Mathematical Economics* 14: 285–299.

Geanokoplos, J., and H. Polemarchakis. 1986. Existence, regularity, and constrained, suboptimality of competitive allocations when markets are incomplete. In *Essays in honor of Kenneth Arrow*, ed. W. Heller, R. Starr, and

D. Starrett, vol. 3. Cambridge: Cambridge University Press.
Geanokoplos, J., and H. Polemarchakis. 1990. Observability and optimality. *Journal of Mathematical Economics* 19: 153–166.
Geanokoplos, J., M. Magill, M. Quinzii, and J. Dreze. 1990. Generic inefficiency of stock market equilibrium when markets are incomplete. *Journal of Mathematical Economics* 19: 113–142.
Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11: 418–443.
Kreps, D. 1979. Three essays on capital markets. Technical Report No. 298. Institute for Mathematical Studies in the Social Sciences, Stanford University.
Radner, R. 1972. Existence of equilibrium of plans, prices, and price expectations. *Econometrica* 40: 289–303.
Townsend, R. 1978. On the optimality of forward markets. *American Economic Review* 68: 54–66.

Increasing Returns to Scale

Spyros Vassilakis

The focus of this essay is the set of positive propositions that can be obtained when technology exhibits increasing returns to scale. The basic incompatibility of perfect competition and increasing returns to scale is examined separately in a section on existence of equilibria, in which we discuss how one should model economies exhibiting such technologies, i.e. essentially how to modify the Walrasian equilibrium concept in order to guarantee existence of equilibria. Welfare and purely empirical problems are not considered.
Definitions: A technology exhibits increasing returns to scale if a proportionate increase in all inputs allows for a more than proportionate increase in outputs; in the single-output case, this implies a decreasing average cost curve.

Division of Labour and Increasing Returns to Scale

Adam Smith (1776), Babbage (1832), Marshall (1890, 1919) and Young (1928) considered the process of division of labour as the main reason why we observe technologies that exhibit increasing returns to scale. A version of their arguments runs as follows: Let A be the set of tasks to be executed in order to produce good x; a partition $A_1,\ldots, A_n$ is called a first-stage division of labour. Each sub-task A_i, $i = 1,\ldots, n$ is executed by (potentially but not necessarily) different kinds of machinery and primary factors, to be called first-stage intermediate goods. The set of tasks to be executed in order to produce each first-stage intermediate good is also subject to division of labour, to be called second-stage division of labour. Each subtask generated by a second-stage division of labour is executed by intermediate goods, to be called second-stage intermediate goods. Clearly, this process can go on indefinitely. We say that the process of division of labour stops at the nth stage if the n-stage intermediate products are all primary factors; a process is feasible if it stops after a finite number of stages and if the demand for primary factors that it generates does not exceed supply. Suppose, now, that production processes are indivisible, i.e. that when an intermediate good is utilized in the production of some other good, its quantity cannot fall short of a minimum irreducible amount, to be called a fixed cost. An increase in the degree of division of labour is defined as either a finer partition of the set of tasks to be executed in order to produce some good, with the number of stages fixed; or an increase in the number of stages. Clearly, then, an increase in the degree of division of labour implies an increase in fixed costs; discarding inferior divisions of labour, therefore, means that an increase in the degree of division of labour has to imply a decrease in variable cost coefficients.

Smith (1776, p. 7) gave three reasons for such a decrease:

> first, … the increase of dexterity in every particular workman; secondly …, the saving of the time which is commonly lost in passing from one species of work to another; and lastly … the invention of a great number of machines which facilitate and abridge labour, and enable one man to do the work of many.

(See also Babbage 1832, ch. xix.) From now on, the (degree of) division of labour and the degree of increasing returns are used as synonyms.

Adam Smith

Adam Smith formulated the following propositions:

(1) The division of labour is limited by the extent of the market (Book 1, ch. 3).
(2) The extent of the market is positively related to population size and density, the amount of natural resources and accumulated capital available, and the ease of transportation (Book 1, ch. 3; Book 2, pp. 259–61).
(3) Small economies devote most of their resources to agriculture, while large economies specialize in industry, because the latter affords a greater degree of division of labour. For exactly the same reason, increases in market size decrease the price of industrial products relative to primary products, and as a consequence the profit rate in industry declines (Book I, ch. XI, pp. 242–7; Book III, ch. I).
(4) Trade increases market size and allows each trader (country, region, individual) to specialize and reap the benefits of increased division of labour. Trade is therefore beneficial to all parties involved, it increases real income of all classes, and therefore should not be restricted by governments (Book IV, ch. II; Book I, ch. II).
(5) Economic activity is located in areas in which transportation is least costly, and therefore in areas with the largest potential for division of labour and trade (Book I, pp. 18–21).
(6) The division of labour is limited by the stability of the market (this is not explicitly stated by Smith, but a number of passages indicated that he was aware of it: Book I, p. 21; Book IV, p. 430).

Notice that (1), (2) and (6) are general propositions, while (3), (4) and (5) are applications.

Proposition (1) has generated many important subsidiary propositions, to be described below. Smith used it to derive (3), (4) and (5) without paying attention to the fact that he never demonstrated how the division of labour is determined (as opposed to limited) by the extent of the market.

Marx (1867, Vol. I, Part IV, section 4), Young (1928), Coase (1937) and Stigler (1951) utilized Proposition (1), often unwittingly, to provide the rudiments of a theory of vertical integration and production roundaboutness. Marx (1867, Vol. I) considered the two as different aspects of the same problem, i.e. vertical (dis)integration is 'division of labour in the society' and production roundaboutness is 'division of labour in the workshop'. The following quotation is from Book I, ch. XIV, section 4, p. 355:

> But what is it that forms the bond between the independent labours of the cattle-breeder, the tanner, and the shoemaker? It is the fact that their respective products are commodities. What, on the other hand, characterizes division of labour in manufactures? The fact that the detail labourer produces no commodities. It is only the common product of all the detail labourers that becomes a commodity. Division of labour in society is brought about by the purchase and sale of the products of different branches of industry, while the connexion between the detail operations in a workshop is due to the sale of labour-power of several workmen in one capitalist, who applies it as combined labour-power. The division of labour in the workshop implies concentration of the means of production in the hands of one capitalist; the division of labour in society implies their dispersion among many independent producers of commodities. While within the workshop, the iron law of proportionality subjects definite numbers of workers to definite functions, in the society outside the workshop, chance and caprice have full play in distributing the producers and their means of production among the various branches of industry.

Marx also saw that the degree of vertical integration is higher the higher the degree of market imperfection:

> the distinction between division of labour in society and in manufacture was practically illustrated to the Yankees. One of the new taxes devised at Washington during the Civil War, was the duty of 6% 'on all industrial products'. Question: What is an industrial product? Answer of the legislature: A thing is produced 'when it is made', and it is made when it is ready for sale . . ..The New York and Philadelphia manufacturers had previously been in the habit of 'making' umbrellas with all their belongings. But since an umbrella is a mixtum compositum of very heterogeneous parts, by degrees these parts became the products of various separate industries, carried on independently in different places. They entered as separate commodities into the umbrella

> manufactory, where they were fitted together. The Yankees have given to articles thus fitted together the name of 'assembled articles', a name they deserve, for being an assemblage of taxes. Thus the umbrella 'assembles' first, 6% on the price of each of its elements, and a further 6% on its own total price' (ibid., p. 355, footnote 2).

Coase (1937) rediscovered and generalized these observations of Marx and constructed a theory of the firm out of them: price-mediated transactions are costly, and firms exist in order to economize on these costs by organizing transactions in a different, non-price mediated way. At this level of generality, the theory is tautological. Stigler (1951) was the first to try to make it operational: he assumes that a single-output firm executes a set of functions, some of them subject to diminishing and others to increasing average cost. The reason why the firm does not become a monopoly is that the increasing cost functions eventually prevail over the decreasing cost ones, so that the firm's average cost curve is U-shaped. (This is clearly not in the spirit of the classical economists, who assumed global increasing returns.) The reason why with small market size, a firm performs the increasing returns functions itself, instead of abandoning them to specialized firms and so sharing fixed costs with other buyers, is that the fixed cost of these functions is too high relative to market size to allow for the survival of even one specialized firm. This argument is based on the implicit assumption that it is profitable for an integrated firm to perform the increasing returns to scale function, while a specialized firm would make a loss because it would not be able to capture all the surplus of the downstream firms, i.e. it would be able to practise only a sufficiently imperfect degree of price discrimination. As market size increases, though, the position of the specialized firm is strengthened, and eventually it can extract enough surplus from the downstream firms to make positive profit; at this point integrated firms abandon the increasing returns function and become downstream firms (buyers) as far as this function is concerned. Spence and Porter (1977) have provided a formal, partial-equilibrium model along these lines.

The nature of the trade-off is different in Vassilakis (1986b): there are global increasing returns to scale, and firms can choose both the degree of division of labour in the production of the final good (i.e. production roundaboutness) and the extert to which they will make their own intermediate goods (vertical integration). Integrated firms do not buy their intermediate goods, and so they avoid monopolistic exploitation associated with the non-price taking behaviour of intermediate goods sellers; on the other hand, they have to pay the fixed cost of producing intermediate goods. For specialist firms the trade-off is reversed. Also, a firm that adopts a high degree of division of labour has to pay higher fixed cost, but lower variable cost, than a firm that produces the same product with a lower degree of division of labour. In equilibrium, the ratio of specialist to integrated firms (the degree of vertical disintegration), and the degree of division of labour within each firm (production roundaboutness), are such that the costs and benefits of marginal changes cancel out. Increases in market size (the number of agents) increase vertical disintegration and production roundaboutness for the same reason: it pays to exploit economies of scale more fully now both by sharing fixed costs with other buyers instead of bearing them unilaterally, and by reducing variable cost through increases in fixed cost. In this sense, market size determines the degree of division of labour. Very clear anticipations of these views on vertical integration are to be found in Austin Robinson (1931, pp. 19, 65, 96, 110).

Proposition (3), another application of Proposition (1), has not been subject to equally intensive theoretical investigation. Kaldar (1978, Essay 9) and Negishi (1986, ch. 3) provide some clarifications. Proposition (4) reappears in Ohlin (1933, ch. 3). For the empirical puzzles that led to the reintroduction of increasing returns to scale in formal trade theory, see Helpman and Krugman (1985, pp. 2–4).

Proposition (5) can be found in Ohlin (1933, pp. 200–211), who generalizes it considerably; increasing returns to scale in production and transportation favour concentration of economic activity in as few points as possible, while the

I

dispersion of natural resources and the fact that certain economics activities are resource-intensive favour decentralization. It is also important whether raw materials for final products are cheaper to transport, with the obvious implications for localization of activities. The result of these consicierations is a generalization of Proposition (5).

> [5] Districts with good transport relations tend to attract pierty of labour and capital and become important market; consequently they tend to specialize in industries which (1) are market-localized and show important advantages from large-scale production; and (2) produce goods which are difficult to transport. On the other hand, districts with poor transport relations become scantily populated and tend to specialize in goods which are easy to transport and can be advantageously produced on a small scale (Ohlin 1933, p. 208).

Implience in Ohlin et al. (1976, pp. 48–50) is the proposition that increases in market size increase geographical concentration of economic activity; the reason seems to be that with increased size there is more to be gained by fuller exploitation of scale economies, i.e. by higher concentration of economic activity, and this gain more than compensates for loss due to increased transportation costs.

Proposition (6) has been exploited by Piore and Sabel (1984). Given that a reduction in demand uncertainty is equivalent to an increase in market size, reductions in uncertainty will increase the degree of division of labour. Piore and Sabel view the coexistence of large and small firms, inventory holding, long-term contracts tying buyers to sellers and vertical integration as uncertainty-reducing devices that allow for a higher degree of division of labour. Also, collective wage bargaining and government stabilization policies are attempts to control that part of uncertainty that cannot be affected by individual firms. Weitzman (1982) and then Kaldor (1983) went even further and argued that a necessary condition for involuntary unemployment, and therefore for Keynesian economics, is the presence of increasing returns to scale, otherwise the unemployed can 'produce themselves out of unemployment', since non-increasing returns to scale imply that small-scale production is at least as efficient as large-scale. (see also the Symposium on Increasing Returns and Unemployment Theory 1985).

Mill and Marx

Mill and Marx gave two closely interrelated propositions:

(7) Increases in market size result in increased concentration of economic activity, in the sense that a higher percentage of the population earn income by selling labour (and not by producing). See Mill (1848, Book I, ch. IX, p. 3) and Marx (1867, ch. XXXII).

(8a) Increases in market size, and the resulting concentration and increase in the scale of production of each firm, is an unqualified benefit from the efficiency point of view, but not necessarily from the equity point of view (Mill ibid.; Marx ibid. and ch. XXV).

Both (7) and (8a) are derived as a consequence of the fact that concentration allows for fuller exploitation of scale economies; Marx added another reason, i.e. that the skills of small-scale producers are 'rendered worthless' by division of labour, which subdivides and simplifies the tasks to be executed in order to produce a commodity (Marx 1848, in McLellan 1977, p. 227).

Another proposition of Marx on the same subject is:

(8b) Increases in market size increase the distance between the economy's actual and potential performance (Marx 1867, ch. XXXII); Elster (1985, ch. 5), provides a rather exhaustive discussion of the exact meaning of this proposition.

We now make (8b) more precise by thinking of increases in market size as generating two contradictory forces: on the one hand, efficiency increases because the increase in market size and the resulting increase in concentration (Proposition 7) allow for fuller exploitation of scale economies; on the other hand, this very

increase in concentration that results in fuller exploitation of scale economies, hampers efficiency by increasing the distortionary effects associated with non-price taking behaviour. In other words, economies of scale are created faster than they are exploited. Finally, we can safely attribute to Marx the following proposition, a variant of his law of the falling rate of profit (Marx 1894, Vol. III, Part III).

(9) Increases in market size reduce the profit rate.

Proposition (9) differs from Proposition (3) of Smith (and Ricardo), because it does not rely on the law of diminishing returns due to land scarcity. (In Marx's words, Ricardo 'fled from economics to seek refuge in organic chemistry' in order to generate a falling profit rate). It is formulated in this particular way, because it has been shown that under constant returns to scale and a constant real wage, the law does not hold, while with a rising real wage it holds only under very restrictive assumptions that turn the law into an improbable special case (Roemer 1981, chs 4, 5 and 6). On the other hand, Negishi (1985, ch. 4) has provided some textual evidence to support the view that Marx had in mind an economy with increasing returns to scale technology and producers facing downward sloping demand, so that (9) is the only version of the law that might be sustainable. Indeed, increasing concentration and a falling profit rate have been obtained in Vassilakis (1986a) as a result of increases in market size; the profit rate, though, falls because both the real wage and the proportion of workers in the population rise in a full employment model, so this version of the law is not entirely in the Marxian spirit. As for Proposition (8a), the formal literature supports the view that with increasing returns to scale only in the neighbourhood of the origin, increases in market size reduce Pareto inefficiency and in the limit they eliminate it (Novshek and Sonnenschein 1978; Hart 1979). On the other hand, Hart and Guesnerie (1985) have found that with global increasing returns, Pareto inefficiency does not disappear in the limit, although per capita inefficiency does; Vassilakis (1986a) finds that even per capita welfare loss can be positive in the limit, for a particular choice of technology; the difference in the result is due to the fact that the latter reference assumes that the alternative to producing is being a worker and earning wage income, while Hart and Guesnerie assume the opportunity cost of a producer to be zero. So, it is fair to say that there is some support for Proposition (8a), while Proposition (8b) remains untested.

Marshall

Marshall (1890, p. 318, 1919, pp. 186–9) believed that all industries exhibit global increasing returns to scale, checked only by short-run fixities or land scarcities; in this case he agreed with the classical economists. As Stigler (1941, p. 78) remarked, though, 'if the economies of large scale production are so important . . ., how do small concerns manage to exist at all?' and 'either the division of labour is limited by the extent of the market, and, characteristically, industries are monopolized; or industries are characteristically competitive, and the theorem is false or of little significance' (Stigler 1951). Marshall tried to reconcile economies of scale and perfect competition in three different ways, namely:

(a) (Some) economies of scale are external to the firm.
(b) Increasing returns to scale is a dynamic phenomenon, and its full effects take so long to manifest themselves that 'the guidance of the business falls into the hands of people with less energy and less creative genius' (Marshall 1890, p. 316).
(c) Transportation costs rise so fast in some industries as to restrict the market area of each firm.

Clearly, (a) assumes the problem away; in Marshall's own words,

> . . . with the growth of capital, the development of machinery, and the improvement of the means of communication the importance of internal economies has increased steadily and fast, while some of the old external economies have declined in importance (Marshall 1919, p. 167).

But even if we assume that most economies of scale are external to the firm, competition is not the most likely outcome; one still has to explain why firms do not merge to internalize external economies, in which case oligopoly is the most likely outcome, or why markets for external effects do not emerge, in which case again, external economies become internal, and we are back to square one. (Starret and Heller (1976) analyse external effects as absence of markets; Makowski (1980) analyses mergers as a way to internalize external effects.)

Explanation (b) is at best of limited importance, unless one can show that expansion by merger is impossible or that the market for managers is so imperfect that a long-lived firm is doomed to fall in the hands of the inept. Also, in Stigler's words, 'if Marshall's discussion of economies is correct and approximately complete, it would not require an extraordinarily high calibre of entrepreneurship to secure a monopoly, or at least a dominant position, in almost any industry' (Stigler 1941, p. 81). Finally, explanation (c) is of limited applicability because it ignores increasing returns to scale in transportation. Marshall himself thought that it cannot be elevated to a general explanation of the coexistence of competition and increasing returns, so he had to invent explanation (b); (Marshall 1919, pp. 315–16). We have to conclude that (not only) in 'competitive, stationary economies, Marshall clearly fails to provide the conditions of stable equilibrium' (Stigler 1941, p. 81). Downward sloping demand and non-price taking behaviour cannot be avoided, therefore; based on Marshall's cues (Marshall 1890, pp. 286–7, 453–8), Sraffa (1926), Robinson (1933) and Chamberlin (1933) reintroduced downward sloping demand almost one hundred years after Cournot.

Despite the fact that Marshall did not have a formal theory of increasing returns economies, he relentlessly applied 'the principle of Increasing Return' to generate propositions. He is the only one after Smith, Marx and Mill to propose a new general proposition (not an application), namely:

(10) '... almost every kind of horizontal extension tends to increase the internal economies of production on a large scale, but as rule, an increase in the variety of output lessens the gain in this direction' (Marshall 1919, p. 216).

In other words, increasing variety reduces efficiency. Proposition (10) is then utilized by Marshall to explain the coexistence of large and small firms, and to determine the range of products of a multiproduct firm. Large firms produce those goods that are most in demand and/or afford the greatest degree of division of labour; their product range is determined by the condition that the addition of one more product would increase cost (due to lost scale economies) by more than it would increase revenue (due to increased market area). Small firms produce goods whose demand is so low, and/or afford so small a degree of division of labour, that large firms do not want to produce, because they can be better off devoting their resources to increase production of the commodities they already produce. As market size increases, there is more to be gained by concentration, so firm size tends to increase. On the other hand, though, small firms will survive at all market sizes, because of three factors (Marshall 1919, ch. III and IV).

(a) The increased income generated by increased market size allows consumers to demand goods closer to their ideal specifications, so the variety of goods demanded increases.
(b) Increased market size increases household specialization, i.e. goods previously produced within the household become commodities.
(c) Increased size increases vertical disintegration.

An obvious implication of this theory is that increases in market size will have different effects, depending on the degree of demand homogeneity and on whether demand is concentrated on goods that afford considerable division of labour. Marshall (1919, Book I) attributes the different growth patterns of industrial economies to differences in

the size stability and perfection of their respective markets.

Existence of Equilibrium

The incompatibility of pricetaking behaviour and increasing returns to scale was first noticed by Cournot (1838, pp. 59–60), but rigorous examination of the issue has been taken up only very recently.

No general existence theorems are available because there is no generally accepted model to imperfect competition. What is available, though, points to the importance of the following three factors: (i) downward sloping demand; (ii) a variable number of firms; (iii) a large number of agents relative to the degree of increasing returns. Downward sloping demand is clearly a necessary condition for existence in economies with global increasing returns to scale, for otherwise firms would have an incentive to produce an unlimited amount of some output. The number of firms should be variable for three reasons: first, because of fixed costs, the number of firms cannot be too large for otherwise profit would be negative; secondly, the number of firms should be sufficiently large to ensure that the demand price faced by each firm is lower than average costs for large enough output levels, otherwise firms would produce arbitrarily large amounts; thirdly, the number of firms should be sufficiently large to discourage entry, so as to ensure that if one more agent sets up a firm, he will earn less than his earnings in the best alternative occupation. Finally, one needs a large number of agents relative to the degree of increasing returns to ensure that the number of firms is large enough to satisfy the conditions above, and in order to convexify reaction correspondences, so that fixed-point theorems can be applied (see Roberts and Sonnenschein 1977, and Novshek and Sonnenschein 1978). All models in the literature on increasing returns to scale base their existence results on (i), (ii) and (iii) above, although they differ in specifics. Thus, we have Bertrand models, in which the agents' strategic variable is price, and Cournot models, in which agents compete in quantities. Also, we have symmetric models, in which all agents are allowed the same strategic possibilities; and non-symmetric models, in which the set of agents is, *a priori* and once and for all, divided into two disjoint sets: the set of consumers–factor suppliers–price takers, and the set of producers–factor demanders–price makers (or quantity setters).

Existence in non-symmetric Cournot models with increasing returns only in a small neighbourhood of the origin is proved in Novshek and Sonnenschein (1978), and with global increasing returns in Hart and Guesnerie (1985); existence in symmetric Cournot models with global increasing returns is proved in Vassilakis (1986a). All proofs with global increasing returns refer to a single-input, single-output economy. For non-symmetric Bertrand games, with a single input, see Hart (1985) and Economides (1982, 1983); for the single-output many-inputs case see Sharkey (1982, ch. 8).

See Also

- ▶ Competition
- ▶ Division of Labour
- ▶ Learning-by-Doing

Bibliography

Babbage, C. 1832. *On the economy of machinery and manufactures*. London: C. Knight.

Chamberlin, E. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Coase, R. 1937. The nature of the firm. *Economica* 4: 386–405.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richeses*. Paris: M. Rivière.

Economides, N. 1982. Oligopoly in differentiated products with three or more competitors. Columbia University discussion paper no. 153.

Economides, N. 1983. Symmetric equilibrium existence and optimality in differentiated products markets. Columbia University discussion paper no. 197.

Elster, J. 1985. *Making sense of Marx*. Cambridge: Cambridge University Press.

Hart, O. 1979. Monopolistic competition in a large economy with differentiated commodities. *Review of Economic Studies* 46: 1–30.
Hart, O. 1985. Monopolistic competition in the spirit of Chamberlin: A general model. *Review of Economic Studies* 52(4): 529–546.
Hart, O., and R. Guesnerie. 1985. Welfare loss due to imperfect competition: Asymptotic results for Cournot-Nash equilibria with and without free entry. *International Economic Review* 26(3): 525–545.
Heller, W., and D. Starret. 1976. On the nature of externalities. In *Theory and measurement of economic externalities*, ed. S. Lin. New York: Academic.
Helpman, E., and P. Krugman. 1985. *Market structure and foreign trade*. Cambridge, MA: MIT Press.
Kaldor, N. 1978. *Further essays on economic theory*. London: Duckworth.
Kaldor, N. 1983. Keynesian economics after fifty years. In *Keynes and the modern world*, ed. D. Worswick and J. Trevithick. Cambridge: Cambridge University Press.
Makowski, L. 1980. Perfect competition, the profit criterion, and the organization of economic activity. *Journal of Economic Theory* 22(2): 222–242.
Marshall, A. 1890. *Principles of economics*. London: Macmillan.
Marshall, A. 1919. *Industry and trade*. London: Macmillan.
Marx, K. 1867–1894. *Capital*, 3 vols. Harmondsworth: Penguin Books, 1976.
McLellan, D. 1977. *Karl Marx, selected writings*. Oxford: Oxford University Press.
Mill, J.S. 1848. *Principles of political economy*. London: J.W. Parker.
Negishi, T. 1985. *Economic theories in a non-Walrasian tradition*. Cambridge: Cambridge University Press.
Novshek, W., and H. Sonnenschein. 1978. Cournot and Walras equilibrium. *Journal of Economic Theory* 19: 223–266.
Ohlin, B. 1933. *Interregional and international trade*. Cambridge, MA: Harvard University Press.
Ohlin, B., et al. 1976. *The international allocation of economic activity*. New York: Holmes & Meier.
Piore, M., and C. Sabel. 1984. *The second industrial divide*. New York: Basic Books.
Roberts, J., and H. Sonnenschein. 1977. On the foundations of the theory of monopolistic competition. *Econometrica* 45: 101–113.
Robinson, E.A.G. 1931. *The structure of competitive industry*. Cambridge: Cambridge University Press.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Roemer, J. 1981. *Analytical foundations of Marxian economic theory*. Cambridge: Cambridge University Press.
Sharkey, W.W. 1982. *The theory of natural monopoly*. Cambridge: Cambridge University Press.
Smith, A. 1776. In *An inquiry into the nature and the courses of the wealth of nations*, ed. E. Cannan. London: Methuen, 1961.
Spence, M., and M. Porter. 1977. Vertical integration and differentiated inputs. Harvard discussion paper no. 576.
Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36: 535–550.
Stigler, G. 1941. *Production and distribution theories*. New York: Macmillan.
Stigler, G. 1951. The division of labor is limited by the extent of the market. *Journal of Political Economy* 59: 185–193.
Symposium on Increasing Returns and Unemployment Theory. 1985. *Journal of Post Keynesian Economics* 7(3): 350–409.
Vassilakis, S. 1986a. Increasing returns and strategic behavior I: The worker-employer ratio. Johns Hopkins working paper no. 168.
Vassilakis, S. 1986b. Increasing returns and strategic behavior II: The division of labor. Johns Hopkins working paper no. 169.
Weitzman, M. 1982. Increasing returns and the foundations of unemployment theory. *Economic Journal* 92: 787–804.
Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.

Indentured Servitude

Farley Grubb

Keywords

Auctions; Colonialism; Incomplete contracts; Indentured servitude; International migration; Labour contracts; Redemption; Slavery

JEL Classifications

N30; N31

Indentured servants were workers – mostly unmarried young adult males – who voluntarily entered alienable forward-labour contracts typically lasting between three and five years in exchange for passage to an overseas destination.

Indentured servitude was important to European overseas expansion and labour migration from the 17th into the 20th century. It was initially prominent among English, Scots, and Irish workers moving to colonies in British America. French and German servants joined this trade in the 18th century, going primarily to Canada and Pennsylvania, respectively (Emmer 1986).

The servant trade had disappeared among British, Irish, and French migrants by the Napoleonic era and among Germans by 1820 (Grubb 1994). Approximately half of the transatlantic migrants in this period were indentured. Servants dominated the colonial labour force early on but, by 1700, African slaves south of Pennsylvania and colonial-born free workers north of Virginia eclipsed them in importance (Galenson 1981; Grubb and Stitt 1994).

After 1830, the repression of the African slave trade and the abolition of slavery in many European colonies led to the revival of the servant trade, especially to tropical sugar-plantation colonies. Between 1834 and 1918 around 1,500,000 indentured servants from India, 250,000 from China, 80,000 from Japan, 50,000 from Portuguese Atlantic islands, and 100,000 from Melanesia were sent to British, French, Dutch, Spanish, German, and US colonies in the Caribbean, Indian Ocean, South and West Africa, Malaya, Australia, Peru, Hawaii, Fiji, and Samoa (Emmer 1986; Northrup 1995).

Servant contracts in the transatlantic trade were typically preprinted single-page forms with blank spaces where negotiated terms were handwritten in. Contracts specified the destination, length of servitude, transferability rights, and 'freedom dues' to be paid at the contract's completion – typically two suits of clothing. In the post-1830 trade freedom dues typically were return-passage tickets. The work to be performed and the maintenance to be received by servants during their contracts were incompletely specified, with contracts typically stating only that servants were to perform customary labour and masters were to provide food, apparel, and lodging (Grubb 2000).

Because passage was provided first, servants had an incentive after arrival to run away or not work hard. Running away was criminalized and runaways were harshly penalized with whippings and forced contract extensions. The disincentive to work was remedied through the contract's incompleteness. With the servants' daily provisions incompletely specified contractually, masters could adjust daily provisions to elicit the optimal daily diligence from servants. Freedom dues compensated servants for their masters' incentive to withhold semi-durable provisions (clothing) from servants near the end of the contract (Grubb 2000).

In the transatlantic trade, markets were largely unregulated and competitive. Servants bargained with shippers over the length of servitude and fixed contract terms before sailing. At debarkation shippers sold these contracts to the highest bidders, thereby recouping their shipping expenses. Competition led to servants signing the shortest contracts necessary to secure passage and to shippers earning zero economic profits on servant cargo. Passage costs were relatively constant across servants but labour productivity was not. Less productive servants had to sign longer contracts for the same passage cost. Contract lengths were inversely related to, whereas auction prices in America were unrelated to, servant productivity known at embarkation (Galenson 1981; Grubb 1985). Servants were also charged about 15 per cent more than free passengers (who paid cash in advance) to compensate shippers for forgoing other investment opportunities and to cover expected servant defaults through mortality, morbidity, and escape.

In the mid-18th century a new variant – redemption – came into use primarily among German immigrants. Under redemption passengers entered fixed-debt passage contracts before sailing that required them to enter servitude at debarkation, if necessary, to clear the debt. Redemption shifted the voyage risk and forecast error in the market from shipper to migrant. With passage debts, but not contract lengths, fixed before sailing, shippers no longer had to forecast at embarkation the amount of labour needed in a servant contract for it to sell at debarkation for enough to cover shipping costs. Instead, at debarkation migrants had to offer however much labour was needed to clear the passage debt contractually guaranteed to the shipper before sailing (Galenson 1981). Migrants accepted this risk because it gave them greater flexibility over selecting their American masters, negotiating contingency clauses into their contracts, and using a single labour contract to pay both the passage debt and any pre-voyage debts transferred to the shipper.

In the post-1830 trade, markets were more highly regulated. For example, in the Melanesian trade to Queensland, Australia, the British government fixed the length of labour contracts at three years and servant wages at six pounds sterling per year, did not allow unrestricted recruiting, and did not allow servants to be auctioned upon arrival. Shippers were licensed to recruit only the number of servants requested by planters and were paid a set fee per recruit. Officials assigned arriving servants to planters according to the number requested. This perversely induced shippers to recruit low-quality labour.

The transatlantic servant trade ended because the supply of servants collapsed, not because American demand declined. Prospective servants found better jobs elsewhere, such as military service during the Napoleonic Wars, or better ways to pay for passage, such as borrowing from already emigrated family members (Grubb 1994). Many post-1830 servant trades were ended by government action or the changing fortunes of the global sugar industry (Emmer 1986; Northrup 1995).

See Also

- ▶ Auctions (Empirics)
- ▶ Compensating Differentials
- ▶ Convict Labour
- ▶ Human Capital, Fertility and Growth
- ▶ International Migration
- ▶ Labour Market Institutions

Bibliography

Emmer, P., ed. 1986. *Colonialism and migration; indentured labour before and after slavery.* Dordrecht: Martinus Nijhoff.

Galenson, D. 1981. *White servitude in colonial America*. New York/Cambridge: Cambridge University Press.

Grubb, F. 1985. The market for indentured immigrants: Evidence on the efficiency of forwardcontracting in Philadelphia, 1745–1773. *Journal of Economic History* 45: 855–868.

Grubb, F. 1994. The end of European immigrant servitude in the United States: An economic analysis of market collapse, 1772–1835. *Journal of Economic History* 54: 794–824.

Grubb, F. 2000. The statutory regulation of colonial servitude: An incomplete-contract approach. *Explorations in Economic History* 37: 42–75.

Grubb, F., and T. Stitt. 1994. The Liverpool emigrant servant trade and the transition to slave labor in the Chesapeake, 1697–1707: Market adjustments to war. *Explorations in Economic History* 31: 376–405.

Northrup, D. 1995. *Indentured labor in the age of imperialism, 1834–1922*. New York/Cambridge: Cambridge University Press.

Index Numbers

W. Erwin Diewert

Abstract

Index numbers are used to aggregate detailed information on prices and quantities into scalar measures of price and quantity levels or their growth. The article reviews four main approaches to bilateral index number theory where two price and quantity vectors are to be aggregated: fixed basket and average of fixed baskets, stochastic, test or axiomatic and economic approaches. The article also considers multilateral index number theory where it is necessary to construct price and quantity aggregates for more than two value aggregates. A final section notes some of the recent literature on related aspects of index number theory.

Keywords

Allen quantity index; Bowley, A.L.; Carli price index; Chain indexes; Consumer price index; Edgeworth, F. Y.; Fisher ideal index; Fisher, I.; Fixed base indexes; Frisch, R. A. K.; Ideal indexes; Index number theory; Index numbers; Jevons price index; Jevons, W. S.; Konüs price index; Konüs-Pollak quantity index; Laspeyres price index; Laspeyres-Konüs quantity index; Logarithmic price ratios; Lowe index; Malmquist quantity index; Marshall, A.; Marshall-Edgeworth index; Paasche price index; Pierson, N. G.; Producer price index; Productivity indexes; Scrope, G. P.; Sidgwick, H.; Sidgwick-Bowley index; Superlative

indexes; Theil, H.; Törnqvist-Theil price index; Walsh index; Walsh, C. M.; Young index

JEL Classifications
C43

Introduction

Each individual consumes the services of thousands of commodities over a year and most producers utilize and produce thousands of individual products and services.

Index numbers are used to reduce and summarize this overwhelming abundance of microeconomic information. Hence index numbers impinge on virtually every empirical investigation in economics.

The index number problem may be stated as follows. Suppose we have price data $p^t \equiv (p_1^t, \ldots, p_N^t)$ and quantity data $q^t \equiv (q_1^t, \ldots, q_N^t)$ on N commodities that pertain to the same economic unit at time period t (or to comparable economic units) for $t = 0, 1, 2, \ldots, T$. The *index number problem* is to find $T+1$ numbers P^t and $T+1$ numbers Q^t such that

$$p^t Q^t = p^t \cdot q^t \equiv \sum_{n=1}^{N} p_n^t q_n^t \text{ for } t = 0, 1, \ldots, T. \tag{1}$$

P^t is the *price index* for period t (or unit t) and Q^t is the corresponding *quantity index*. P^t is supposed to be representative of all of the prices $p_n^t, n = 1, \ldots, N$ in some sense, while Q^t is to be similarly representative of the quantities q_n^t, $n = 1, \ldots, N$. In what precise sense P^t and Q^t represent the individual prices and quantities is not immediately evident, and it is this ambiguity that leads to different approaches to index number theory. Note that we require that the product of the price and quantity indexes, $P^t Q^t$, equals the actual period (or unit) t expenditures on the N commodities, $p^t \cdot q^t$. Thus if the P^t are determined, then the Q^t may be implicitly determined using eq. (1), or vice versa.

The number P^t is interpreted as an aggregate period t price level while the number Q^t is interpreted as an aggregate period t quantity level. The *levels* approach to index number theory works as follows. The aggregate price level P^t is assumed to be a function of the components in the period t price vector, p^t while the aggregate period t quantity level Q^t is assumed to be a function of the period t quantity vector components, q^t; that is, it is assumed that

$$P^t = c(p^t) \text{ and } Q^t = f(q^t); \; t = 0, 1, \ldots, T. \tag{2}$$

The functions c and f are to be determined somehow. Note that we are requiring that the functional forms for the price aggregation function c and for the quantity aggregation function f be independent of time. This is a reasonable requirement since there is no reason to change the method of aggregation as time changes.

Substituting (2) into (1) and dropping the superscripts t means that c and f must satisfy the following functional equation for all strictly positive price and quantity vectors:

$$c(p)f(q) = p \cdot q \equiv \sum_{n=1}^{N} p_n q_n \text{ for all } p >> 0_N \text{ and for all } q >> 0_N. \tag{3}$$

Note that $p >> 0_N$ means that each component of p is positive, $p \geq 0_N$ means each component is non-negative and $p > 0_N$ means each component is non-negative and at least one component is positive. We now could ask what properties the price aggregation function c and the quantity aggregation function f should have. We could assume that c and f satisfied various 'reasonable' properties and hope that these properties would determine the functional form for c and f. However, it turns out that we have only to make the following very weak *positivity* assumptions on f and c in order to obtain an impossibility result:

I

$$c(p) > 0 \text{ for all } p >> 0_N; \quad f(q) > 0 \text{ for all } q >> 0_N. \tag{4}$$

Eichhorn (1978, p. 144) proved the following result: if the number of commodities N is greater than 1, then there do not exist any functions c and f that satisfy (3) and (4). Thus this *levels approach* to index number theory comes to an abrupt halt. As we shall see later, when the economic approach to index number theory is studied, this is not quite the end of the story: in (3) and (4), we allowed p and q to vary independently from each other, and this is what leads to the impossibility result. If instead we allow p to vary independently but assume that q is determined as the result of an optimizing model, then eq. (3) can be satisfied.

If we change the question that we are trying to answer slightly, then there are practical solutions to the index number problem. The change is that instead of trying to decompose the value of the aggregate into price and quantity components for a single period, we instead attempt to decompose a *value ratio* pertaining to two periods, say periods 0 and 1, into a *price change component* P times a *quantity change component* Q. Thus we now look for two functions of $4N$ variables, $P(p^0, p^1, q^0, q^1)$ and $Q(p^0, p^1, q^0, q^1)$ so that:

$$p^1.q^1/p^0.q^0 = P\left(p^0,p^1,q^0,q^1\right)\ Q\left(p^0,p^1,q^0,q^1\right). \tag{5}$$

Note that if some approach to index number theory determines the 'best' functional form for the price index $P(p^0, p^1, q^0, q^1)$, then the *product test* (5) can be used to determine the functional form for the corresponding quantity index, $Q(p^0, p^1, q^0, q^1)$.

If we take the *test or axiomatic approach* to index number theory, then we want eq. (5) to hold for all positive price and quantity vectors pertaining to the two periods under consideration, p^0, p^1, q^0, q^1. If we take the *economic approach*, then only the price vectors p^0 and p^1 are regarded as independent variables while the quantity vectors, q^0 and q^1, are regarded as dependent variables. In section "The Test Approach to Index Number Theory" below, we will pursue the test approach and in sections "The Economic Approach to Price Indexes," "Economic Approaches to Quantity Indexes," and "Exact and Superlative Indexes," we will take the economic approach. In sections "Fixed Basket Approaches," "The Stochastic Approach to Index Number Theory," "The Test Approach to Index Number Theory," "The Economic Approach to Price Indexes," "Economic Approaches to Quantity Indexes," and "Exact and Superlative Indexes," we take a *bilateral approach to index number theory*; that is, in making price and quantity comparisons between any two time periods, the relevant indexes use *only* price and quantity information that pertains to the two periods under consideration. It is also possible to take a *multilateral approach*; that is, we look for functions, P^t and Q^t, that are functions of *all* of the price and quantity vectors, $p^0, p^1, \ldots, p^T, q^0, q^1, \ldots, q^T$. Thus we look for $2(T+1)$ functions, $P^t(p^0, p^1, \ldots, p^T, q^0, q^1, \ldots, q^T)$ and $Q^t(p^0, p^1, \ldots, p^T, q^0, q^1, \ldots, q^T)$, $t = 0, 1, \ldots, T$, so that

$$p^t \cdot q^t = P^t\left(p^0,p^1,\ldots,p^T,q^0,q^1,\ldots;q^T\right) \times Q^t\left(p^0,p^1,\ldots,p^T,q^0,q^1,\ldots,q^T\right) \quad \text{for } t = 0, 1, \ldots, T. \tag{6}$$

We briefly pursue the multilateral approach to index number theory in section "Multilateral Indexes".

The four main approaches to bilateral index number theory will be covered in this review: (i) the *fixed basket approach* (section "Fixed Basket Approaches"), (ii) the *stochastic approach* (section "The Stochastic Approach to Index Number Theory"), (iii) the *test approach* (section "The Test Approach to Index Number Theory") and (iv) the *economic approach*, which relies on the assumption of maximizing or minimizing behaviour (sections "The Economic Approach to Price Indexes," "Economic Approaches to Quantity Indexes," and "Exact and Superlative Indexes").

Section "The Fixed Base Versus the Chain Principle" discusses fixed base versus chained index numbers, and section "Other Aspects of Index Number Theory" concludes by mentioning some recent areas of active research in the index number literature.

Fixed Basket Approaches

The English economist Joseph Lowe (1823) developed the theory of the consumer price index in some detail. His approach to measuring the price change between periods 0 and 1 was to specify an approximate representative commodity basket quantity vector, $q \equiv (q_1, \ldots, q_N)$, which was to be updated every 5 years, and then calculate the level of prices in period 1 relative to period 0 as

$$P_{Lo}(p^0,p^1,q)\ p^1 \cdot q/p^0 \cdot q \tag{7}$$

where p^0 and p^1 are the commodity price vectors that the consumer (or group of consumers) face in periods 0 and 1 respectively. The fixed basket approach to measuring price change is intuitively very simple: we simply specify the commodity 'list' q and calculate the price index as the ratio of the costs of buying this same list of goods in periods 1 and 0.

As time passed, economists and price statisticians demanded more precision with respect to the specification of the basket vector q. There are two natural choices for the reference basket: the period 0 commodity vector q^0 or the period 1 commodity vector q^1. These two choices lead to the Laspeyres (1871) price index P_L defined by (8) and the Paasche (1874) price index P_P defined by (9):

$$P_L(p^0,p^1,q^0,q^1) \equiv p^1 \cdot q^0/p^0 \cdot q^0; \tag{8}$$

$$P_p(p^0,p^1,q^0,q^1) \equiv p^1 \cdot q^1/p^0 \cdot q^1. \tag{9}$$

The above formulae can be rewritten in an alternative manner that is very useful for statistical agencies. Define the period t expenditure share on commodity n as follows:

$$s_n^t \equiv p_n^t q_n^t/p^t.q^t \text{ for } n = 1, \ldots, N \text{ and } t = 0, 1. \tag{10}$$

Following Fisher (1911), the Laspeyres index (8) can be rewritten as follows:

$$\begin{aligned} P_L(p^0,p^1,q^0,q^1) &= \sum_{n=1}^{N} p_n^1\ q_n^0/p^0 \cdot q^0 \\ &= \sum_{n=1}^{N} (p_n^1/p_n^0) p_n^0\ q_n^0/p^0 \cdot q^0 \\ &= \sum_{n=1}^{N} (p_n^1/p_n^0) \\ &\quad \times s_n^0 \text{ using definitions (10).} \end{aligned} \tag{11}$$

Thus the Laspeyres price index P_L can be written as a base period expenditure share weighted average of the N price ratios (or price relatives using index number terminology), p_n^1/p_n^0. The Laspeyres formula (until the very recent past when in 2003 the US Bureau of Labor Statistics introduced its chained consumer price index) has been widely used as the intellectual basis for country consumer price indexes (CPIs) around the world. To implement the formula, the country statistical agency collects information on expenditure shares s_n^0 for the index domain of definition for the base period 0 and then collects information on *prices* alone on an ongoing basis. Thus a Laspeyres-type CPI can be produced on a timely basis without one having to know current period quantity information. In fact, the situation is more complicated than this: in actual CPI programmes, prices are collected on a monthly or quarterly frequency and with base month 0 say, but the quantity vector q^0 is typically *not* the quantity vector that pertains to the price base month 0; rather, it is actually equal to a *base year quantity vector*, q^b say, which is typically prior to the base month 0. Thus the typical CPI, although loosely based on the Laspeyres index, is actually a form of Lowe index; see (7) above. Instead of using the Lowe formula for their CPI, some statistical agencies use the following Young (1812) index:

$$P_Y(p^0,p^1,s^b) \equiv \sum_{n=1}^{N} (p_n^1/p_n^0) s_n^b \tag{12}$$

where the s_n^b are base year expenditure shares on the N commodities in the index. For additional

material on Lowe and Young indexes and their use in CPI and producer price index (PPI) programmes, see the ILO (2004) and the IMF (2004).

The Paasche index can also be written in expenditure share and price ratio form as follows:

$$\begin{aligned} P_p(p^0,p^1,q^0,q^1) &= 1/\left[\sum_{n=1}^{N} p_n^0 q_n^1/p^1.q^1\right] \\ &= 1/\left[\sum_{n=1}^{N} (p_n^0/p_n^1) p_n^1 q_n^1/p^1.q^1\right] \\ &= 1/\left[\sum_{n=1}^{N} (p_n^1/p_n^0)^{-1} s_n^1\right] \text{ using definitions} \\ (10) &= \left[\sum_{n=1}^{N} (p_n^1/p_n^0)^{-1} s_n^1\right]^{-1}. \end{aligned} \tag{13}$$

Thus the Paasche price index P_P can be written as a period 1 (or current period) expenditure share weighted *harmonic* average of the N price ratios.

The problem with the Paasche and Laspeyres index number formulae is that they are equally plausible but, in general, they will give different answers. This suggests that, if we require a single estimate for the price change between the two periods, then we need to take some sort of evenly weighted average of the two indexes as our final estimate of price change between periods 0 and 1. Examples of such symmetric averages are the arithmetic mean, which leads to the Sidgwick (1883, p. 68) and Bowley (1901, p. 227) index, $(1/2)P_L + (1/2)P_P$, and the geometric mean, which leads to the Fisher (1922) ideal index, P_F, which was actually first suggested by Bowley (1899, p. 641), defined as

$$\begin{aligned} &P_F(p^0,p^1,q^0,q^1) \\ &\equiv [P_L(p^0,p^1,q^0,q^1) P_P(p^0,p^1,q^0,q^1)]^{1/2}. \end{aligned} \tag{14}$$

At this point, the fixed basket approach to index number theory is transformed into the *test approach* to index number theory; that is, in order to determine which of these fixed basket indexes or which averages of them might be best, we need *criteria* or *tests* or *properties* that we would like our indexes to satisfy. We will pursue this topic in more detail in section "The Test Approach to Index Number Theory," but we give the reader an introduction to this topic in the present section because some of these tests or properties are useful to evaluate other approaches to index number theory.

Let a and b be two positive numbers. Diewert (1993b, p. 361) defined a *symmetric mean* of a and b as a function $m(a, b)$ that has the following properties: (i) $m(a, a) = a$ for all $a > 0$ (mean property); (ii) $m(a, b) = m(b, a)$ for all $a > 0, b > 0$ (symmetry property); (iii) $m(a, b)$ is a continuous function for $a > 0, b > 0$ (continuity property) and (iv) $m(a, b)$ is a strictly increasing function in each of its variables (increasingness property). Eichhorn and Voeller (1976, p. 10) showed that, if $m(a, b)$ satisfies the above properties, then it also satisfies the following property: (v) $\min\{a, b\} \leq m\{a, b\} \leq \max\{a, b\}$ (min-max property); that is, the mean of a and b, $m(a, b)$, lies between the maximum and minimum of the numbers a and b. Since we have restricted the domain of definition of a and b to be positive numbers, it can be seen that an implication of the last property is that m also satisfies the following property: (vi) $m(a, b) > 0$ for all $a > 0$, $b > 0$ (positivity property). If in addition, m satisfies the following property, then we say that m is a *homogeneous symmetric mean*: (vii) $m(\lambda a, \lambda b) = \lambda m(a, b)$ for all $\lambda > 0$, $a > 0$, $b > 0$.

What is the best symmetric average of P_L and P_P to use as a point estimate for the theoretical cost of living index? It is very desirable for a price index formula that depends on the price and quantity vectors pertaining to the two periods under consideration to satisfy the *time reversal test*. We say that the index number formula $P(p^0, p^1, q^0, q^1)$ satisfies this test if

$$P(p^1,p^0,q^1,q^0) = 1/P(p^0,p^1,q^0,q^1); \tag{15}$$

that is, if we interchange the period 0 and period 1 price and quantity data and evaluate the index, then this new index $P(p^1, p^0, q^1, q^0)$ is equal to the reciprocal of the original index $P(p^0, p^1, q^0, q^1)$. For the history of this test (and other tests), see Diewert (1992a, p. 218, 1993a).

Diewert (1997, p. 138) proved the following result: the Fisher ideal price index defined by (14)

above is the *only* index that is a homogeneous symmetric average of the Laspeyres and Paasche price indexes, P_L and P_P, that also satisfies the time reversal test (15) above.

Thus the symmetric basket approach to index number theory leads to the Fisher ideal index as the best formula. It is interesting to note that this symmetric basket approach to index number theory dates back to Bowley, one of the early pioneers of index number theory, as the following quotations indicate:

> If [the Paasche index] and [the Laspeyres index] lie close together there is no further difficulty; if they differ by much they may be regarded as inferior and superior limits of the index number, which may be estimated as their arithmetic mean ... as a first approximation. (Bowley 1901, p. 227)

> When estimating the factor necessary for the correction of a change found in money wages to obtain the change in real wages, statisticians have not been content to follow Method II only [to calculate a Laspeyres price index], but have worked the problem backwards [to calculate a Paasche price index] as well as forwards. ... They have then taken the arithmetic, geometric or harmonic mean of the two numbers so found. (Bowley 1919, p. 348)

Instead of taking a symmetric average of the Paasche and Laspeyres indexes, an alternative average basket approach takes a symmetric average of the baskets that prevail in the two periods under consideration. For example, the average basket could be the arithmetic or geometric mean of the two baskets, leading the Marshall (1887) and Edgeworth (1925) index P_{ME} or the Walsh (1901, p. 398, 1921a, pp. 97–101) index P_W:

$$P_{ME}\left(P^0,P^1,q^0,q^1\right) \equiv \sum_{n=1}^{N} p_n^1 (1/2)\left(q_n^0+q_n^1\right) / \sum_{m=1}^{N} P_j^0 (1/2)\left(q_m^0+q_m^1\right); \tag{16}$$

$$P_W\left(P^0,P^1,q^0,q^1\right) \equiv \sum_{n=1}^{N} p_n^1 \left(q_n^0 q_n^1\right)^{1/2} / \sum_{m=1}^{N} P_m^0 \left(q_m^0,q_m^1\right)^{1/2}. \tag{17}$$

Diewert (2002b, pp. 569–71) showed that the Walsh index P_W emerged as being best in this average basket framework; see also ILO (2004, chs 15 and 16).

We turn now to the second major approach to bilateral index number theory.

The Stochastic Approach to Index Number Theory

> In drawing our averages the independent fluctuations will more or less destroy each other; the one required variation of gold will remain undiminished. (Jevons 1884, p. 26)

The stochastic approach to the determination of the price index can be traced back to the work of Jevons (1865, 1884) and Edgeworth (1888, 1923, 1925) over 100 years ago. For additional discussion on the early history of this approach, see Diewert (1993a, pp. 37–8, 1995b).

The basic idea behind the stochastic approach is that each price relative, p_n^1/p_n^0 for $n = 1,2,\ldots,N$ can be regarded as an estimate of a common inflation rate α between periods 0 and 1; that is, it is assumed that

$$p_n^1/p_n^0 = \alpha + \varepsilon_n; n = 1,2,\ldots,N \tag{18}$$

where α is the common inflation rate and the ε_n are random variables with mean 0 and variance σ^2. The least squares estimator for α is the Carli (1764) price index P_C defined as

$$P_C\left(p^0,p^1\right) \equiv \sum_{n=1}^{N} (1/N)\left(p_n^1/p_n^0\right). \tag{19}$$

Unfortunately, P_C does not satisfy the time reversal test, namely, $P_C(p^1, p^0) \neq 1/P_C(p^0, p^1)$. In fact, Fisher (1922, p. 66) noted that $P_C(p^0, p^1)$ $P_C(p^1, p^0) \geq 1$ unless the period 1 price vector p^1 is proportional to the period 0 price vector $p0$; that is, Fisher showed that the Carli (and the Young) index has a definite upward bias. He urged statistical agencies not to use these formulae.

Now assume that the logarithm of each price relative, $\ln\left(p_n^1/p_n^0\right)$, is an unbiased estimate of the logarithm of the inflation rate between periods 0 and 1, β say. Thus we have:

$$\ln\left(p_n^1/p_n^0\right), = \beta + \varepsilon_n; n = 1, 2, \ldots, N \qquad (20)$$

where $\beta \equiv \ln \alpha$ and the ε_n are independently distributed random variables with mean 0 and variance σ^2. The least squares estimator for β is the logarithm of the geometric mean of the price relatives. Hence the corresponding estimate for the common inflation rate α is the Jevons (1865) price index P_J defined as:

$$P_J\left(p^0, p^1\right) \equiv \prod_{n=1}^{N} \left(p_n^1/p_n^0\right)^{1/N}. \qquad (21)$$

The Jevons price index P_J satisfies the time reversal test and hence is much more satisfactory than the Carli index P_C.

Bowley (1928) attacked the use of both (19) and (21) on two grounds. First, from an empirical point of view, he showed that price ratios were not symmetrically distributed about a common mean and their logarithms also failed to be symmetrically distributed. Second, from a theoretical point of view, he argued that it was unlikely that prices or price ratios were independently distributed. Keynes (1930) developed Bowley's second objection in more detail; he argued that changes in the money supply would not affect all prices at the same time. Moreover, real disturbances in the economy could cause one set of prices to differ in a systematic way from other prices, depending on various elasticities of substitution and complementarity. In other words, prices are not randomly distributed, but are systematically related to each other through the general equilibrium of the economy. Keynes (1930, pp. 76–7) had other criticisms of this *unweighted stochastic approach* to index number theory, including the point that that there is no such thing as *the* inflation rate; there are only price changes that pertain to well-specified sets of commodities or transactions; that is, the domain of definition of the price index must be carefully specified. Keynes also followed Walsh in insisting that price movements must be weighted by their economic importance, that is, by quantities or expenditures:

> It might seem at first sight as if simply every price quotation were a single item, and since every commodity (any kind of commodity) has one price-quotation attached to it, it would seem as if price-variations of every kind of commodity were the single item in question. This is the way the question struck the first inquirers into price-variations, wherefore they used simple averaging with even weighting. But a price-quotation is the quotation of the price of a generic name for many articles; and one such generic name covers a few articles, and another covers many. ... A single price-quotation, therefore, may be the quotation of the price of a hundred, a thousand, or a million dollar's worth, of the articles that make up the commodity named. Its weight in the averaging, therefore, ought to be according to these money-unit's worth. (Walsh 1921a, pp. 82–3)

Theil (1967, pp. 136–7) proposed a solution to the lack of weighting in (21). He argued as follows. Suppose we draw price relatives at random in such a way that each dollar of expenditure in the base period has an equal chance of being selected. Then the probability that we will draw the nth price relative is equal to $s_n^0 \equiv p_n^0 q_n^0 / p^0 . q^0$, the period 0 expenditure share for commodity n. Then the overall mean (period 0 weighted) logarithmic price change is $\sum_{n=1}^{N} s_n^0 \ln \left(p_n^1/p_n^0\right)$. Now repeat the above mental experiment and draw price relatives at random in such a way that each dollar of expenditure in period 1 has an equal probability of being selected. This leads to the overall mean (period 1 weighted) logarithmic price change of $\sum_{n=1}^{N} s_n^1 \ln \left(p_n^1/p_n^0\right)$. Each of these measures of overall logarithmic price change seems equally valid so we could argue for taking a symmetric average of the two measures in order to obtain a final single measure of overall logarithmic price change. Theil (1967, p. 138) argued that a nice symmetric index number formula can be obtained if we make the probability of selection for the nth price relative equal to the arithmetic average of the period 0 and 1 expenditure shares for commodity n. Using these probabilities of selection, Theil's final measure of overall logarithmic price change was

$$\ln P_T(p^0, p^1, q^0, q^1) \equiv \sum_{n=1}^{N} (1/2)\left(s_n^0 + s_n^1\right) \times \ln\left(p_n^1/p_n^0\right). \qquad (22)$$

We can give the following *descriptive statistics* interpretation of the right hand side of (22). Define the *n*th logarithmic price ratio *rn* by:

$$r^n \equiv \ln\left(p_n^1/p_n^0\right) \quad \text{for } n = 1, \ldots, N. \tag{23}$$

Now define the discrete random variable, R say, as the random variable which can take on the values r_n with probabilities $\rho_n \equiv (1/2)\left[s_n^0 + s_n^1\right]$ for $n = 1, \ldots, N$. Note that, since each set of expenditure shares, s_n^0 and s_n^1, sums to one, the probabilities ρ_n will also sum to one. It can be seen that the expected value of the discrete random variable R is

$$\begin{aligned} E[R] &\equiv \sum_{n=1}^{N} \rho_n r_n \\ &= \sum_{n=1}^{N} (1/2)\left(s_n^0 + s_n^1\right) \ln\left(p_n^1/p_n^0\right) \\ &= \ln P_T\left(p^0, p^1, q^0, q^1\right) \end{aligned} \tag{24}$$

using (22) and (23). Thus the logarithm of the index P_T can be interpreted as *the expected value of the distribution of the logarithmic price ratios* in the domain of definition under consideration, where the N discrete price ratios in this domain of definition are weighted according to Theil's probability weights $\rho_n \equiv (1/2)\left[s_n^0 + s_n^1\right]$ for $n = 1, \ldots, N$.

If we take antilogs of both sides of (24), we obtain the Törnqvist (1936) and Törnqvist and Törnqvist (1937). Theil price index, P_T. This index number formula has a number of good properties. Thus the second major approach to bilateral index number theory has led to the Törnqvist-Theil price index P_T as being best from this perspective.

Additional material on stochastic approaches to index number theory and references to the literature can be found in Selvanathan and Rao (1994), Diewert (1995b), Wynne (1997), ILO (2004), IMF (2004), and Clements et al. (2006).

Formulae (8), (9), (14) and (22) (the Laspeyres, Paasche, Fisher and Törnqvist-Theil formulae) are the most widely used formulae for a bilateral price index. But Walsh (1901) and Fisher (1922) presented hundreds of functional forms for bilateral price indexes – on what basis are we to choose one as being better than the other? Perhaps the next approach to index number theory will narrow the choices.

The Test Approach to Index Number Theory

In this section, we will take the perspective outlined in section "Introduction" above; that is, along with the price index $P(p^0, p^1, q^0, q^1)$, there is a companion quantity index $Q(p^0, p^1, q^0, q^1)$ such that the product of these two indexes equals the value ratio between the two periods. Thus, throughout this section, we assume that P and Q satisfy the product test (5) above.

If we assume that the product test holds means that as soon as the functional form for the price index P is determined, then (5) can be used to determine the functional form for the quantity index Q. However, as Fisher (1911, pp. 400–6) and Vogt (1980) observed, a further advantage of assuming that the product test holds is that we can assume that the quantity index Q satisfies a 'reasonable' property and then use (5) to translate this test on the quantity index into a corresponding test on the price index P.

If $N = 1$, so that there is only one price and quantity to be aggregated, then a natural candidate for P is p_1^1/p_1^0, the single price ratio, and a natural candidate for Q Is q_1^1/q_1^0, the single quantity ratio. When the number of commodities or items to be aggregated is greater than 1, then what index number theorists have done over the years is to propose properties or tests that the price index P should satisfy. These properties are generally multidimensional analogues to the one good price index formula, p_1^1/p_1^0. Below, following Diewert (1992a), we list 20 tests that characterize the Fisher ideal price index.

We shall assume that every component of each price and quantity vector is positive; that is, $p^t >> 0_N$ and $q^t >> 0_N$ for $t = 0,1$. If we want to set $q^0 = q^1$, we call the common quantity vector q; if we want to set $p^0 = p^1$, we call the common price vector p.

Our first two tests, due to Eichhorn and Voeller (1976, p. 23) and Fisher (1922, pp. 207–15), are not very controversial and so we will not discuss them.

T1: *Positivity*: $P(p^0, p^1, q^0, q^1) > 0$.
T2: *Continuity*: $P(p^0, p^1, q^0, q^1)$ is a continuous function of its arguments.

Our next two tests, due to Laspeyres (1871, p. 308), Walsh (1901, p. 308), and Eichhorn and Voeller (1976, p. 24), are somewhat more controversial.

T3: *Identity or constant prices test*: $P(p, p, q^0, q^1) = 1$.

That is, if the price of every good is identical during the two periods, then the price index should equal unity, no matter what the quantity vectors are. The controversial part of this test is that the two quantity vectors are allowed to be different in the above test.

T4: *Fixed basket or constant quantities test*: $P(p^0, p^1, q, q) = \sum_{i=1}^{N} p_j^1 q_i / \sum_{i=1}^{N} p_i^0 qi$.

That is, if quantities are constant during the two periods so that $q^0 = q^1 \equiv q$, then the price index should equal the expenditure on the constant basket in period 1, $\sum_{i=1}^{N} p_i^1 qi$, divided by the expenditure on the basket in period 0, $\sum_{i=1}^{N} p_i^0 qi$. The origins of this test go back at least 200 years to the Massachusetts legislature which used a constant basket of goods to index the pay of Massachusetts soldiers fighting in the American Revolution: see Willard Fisher (1913). Other researchers who have suggested the test over the years include Lowe (1823, Appendix, p. 95), Scrope (1833, p. 406), Jevons (1865), Sidgwick (1883, pp. 67–8), Edgeworth (1887, p. 215), Marshall (1887, p. 363), Pierson (1895, p. 332), Walsh (1901, p. 540, 1921b, p. 544), and Bowley (1901, p. 227). Vogt and Barta (1997, p. 49) also observed that this test is a special case of Fisher's (1911, p. 411) proportionality test for quantity indexes which Fisher (1911, p. 405) translated into a test for the price index using the product test (5).

The following four tests restrict the behaviour of the price index P as the scale of any one of the four vectors p^0, p^1, q^0, q^1 changes. The following test was proposed by Walsh (1901, p. 385), Eichhorn and Voeller (1976, p. 24), and Vogt (1980, p. 68).

T5: *Proportionality in Current Prices*: $P(p^0, \lambda p^1, q^0, q^1) = \lambda P(p^0, p^1, q^0, q^1)$ for $\lambda > 0$.

That is, if all period 1 prices are multiplied by the positive number λ, then the new price index is λ times the old price index. Put another way, the price index function $P(p^0, p^1, q^0, q^0)$ is (positively) homogeneous of degree one in the components of the period 1 price vector $p1$. Most index number theorists regard this property as a very fundamental one that the index number formula should satisfy.

Walsh (1901) and Fisher (1911, p. 418, 1922, p. 420) proposed the related proportionality test $P(p, \lambda p, q^0, q^1) = \lambda$. This last test is a combination of T3 and T5; in fact Walsh (1901, p. 385) noted that this last test implies the identity test, T3.

In the next test, due to Eichhorn and Voeller (1976, p. 28), instead of multiplying all period 1 prices by the same number, we multiply all period 0 prices by the number λ.

T6: *Inverse proportionality in base period prices*: $P(\lambda p^0, p^1, q^0, q^1) = \lambda^{-1} P.\ (p^0, p^1, q^0, q^1)$ for $\lambda > 0$.

That is, if all period 0 prices are multiplied by the positive number λ, then the new price index is $1/\lambda$ times the old price index. Put another way, the price index function $P(p^0, p^1, q^0, q^1)$ is (positively) homogeneous of degree minus one in the components of the period 0 price vector p^0.

The following two homogeneity tests can also be regarded as invariance tests.

T7: *Invariance to proportional changes in current quantities*: $P(p^0, p^1, q^0, \lambda q^1) = P(p^0, p^1, q^0, q^1)$ for all $\lambda > 0$.

That is, if current period quantities are all multiplied by the number λ, then the price index

remains unchanged. Put another way, the price index function $P(p^0, p^1, q^0, q^1)$ is (positively) homogeneous of degree zero in the components of the period 1 quantity vector q^1. Vogt (1980, p. 70) was the first to propose this test and his derivation of the test is of some interest. Suppose the quantity index Q satisfies the quantity analogue to the price test T5, that is, suppose Q satisfies $Q(p^0, p^1, q^0, \lambda q^1) = \lambda Q(p^0, p^1, q^0, q^1)$ for $\lambda > 0$. Then using the product test (5), we see that P must satisfy T7.

T8: *Invariance to proportional changes in base quantities*: $P(p^0, p^1, \lambda q^0, q^1) = P(p^0, p^1, q^0, q^1)$ for all $\lambda > 0$.

That is, if base period quantities are all multiplied by the number λ, then the price index remains unchanged. Put another way, the price index function $P(p^0, p^1, q^0, q^1)$ is (positively) homogeneous of degree zero in the components of the period 0 quantity vector q^0. If the quantity index Q satisfies the following counterpart to T8: $Q(p^0, p^1, \lambda q^0, q^1) = \lambda^{-1} Q(p^0, p^1, q^0, q^1)$ for all $\lambda > 0$, then, using (5), the corresponding price index P must satisfy T8. This argument provides some additional justification for assuming the validity of T8 for the price index function P. This test was proposed by Diewert (1992a, p. 216).

T7 and T8 together impose the property that the price index P does not depend on the *absolute* magnitudes of the quantity vectors q^0 and q^1.

The next five tests are invariance or symmetry tests. Fisher (1922, pp. 62–3, 458–60) and Walsh (1921b, p. 542) seem to have been the first researchers to appreciate the significance of these kinds of tests. Fisher (1922, pp. 62–3) spoke of fairness but it is clear that he had symmetry properties in mind. It is perhaps unfortunate that he did not realize that there were more symmetry and invariance properties than the ones he proposed; if he had realized this, it is likely that he would have been able to provide an axiomatic characterization for his ideal price index, as will be done shortly. Our first invariance test is that the price index should remain unchanged if the *ordering* of the commodities is changed:

T9: *Commodity reversal test* (or invariance to changes in the ordering of commodities):

$$P(p^{0*}, p^{1*}, q^{0*}, q^{1*}) = P(p^0, p^1, q^0, q^1)$$

where p^{t*} denotes a permutation of the components of the vector p^t and q^{t*} denotes the same permutation of the components of q^t for $t = 0, 1$. This test is due to Fisher (1922), and it is one of his three famous reversal tests. The other two are the time reversal test and the factor reversal test which will be considered below.

T10: *Invariance to changes in the units of measurement* (commensurability test):

$$P(\alpha_1 p_1^0, \ldots, \alpha_N p_N^0; \alpha_1 p_1^1, \ldots, \alpha_N p_N^1; \alpha_1^{-1} q_1^0, \ldots, \alpha_N^{-1} q_N^0; \alpha_1^{-1} q_1^1, \ldots, \alpha_N^{-1} q_N^1)$$
$$= p(p_1^0, \ldots, p_N^0; p_1^1, \ldots p_N^1; q_1^0, \ldots, q_N^0; q_1^1, \ldots, q_N^1) \text{ for all } \alpha_1 > 0, \ldots, \alpha_N > 0.$$

That is, the price index does not change if the units of measurement for each commodity are changed. The concept of this test was due to Jevons (1884, p. 23) and the Dutch economist Pierson (1896, p. 131), who criticized several index number formula for not satisfying this fundamental test. Fisher (1911, p. 411) first called this test *the change of units test* and later, Fisher (1922, p. 420) called it the *commensurability test*.

T11: *Time reversal test*: $P(p^0, p^1, q^0, q^1) = 1/P(p^1, p^0, q^1, q^0)$.

That is, if the data for periods 0 and 1 are interchanged, then the resulting price index should equal the reciprocal of the original price index. We have already encountered this test: see (15) above. Obviously, in the one good case when the price index is simply the single price ratio, this test is satisfied (as are all of the other tests listed in this section). When the number of goods is greater than one, many commonly used price indexes fail this test; for example, the Laspeyres and Paasche price indexes, P_L and P_P defined earlier by (8) and (9) above, both *fail* this fundamental test. The concept of the test was due to Pierson (1896, p. 128), who was so upset by the fact that many of the commonly used index number formulae did not satisfy this test that he proposed that the entire concept of an index number should be abandoned. More formal statements of the test were made by Walsh (1901, p. 368; 1921b, p. 541) and Fisher (1911, p. 534; 1922, p. 64).

Our next two tests are more controversial, since they are not necessarily consistent with the economic approach to index number theory. However, these tests are quite consistent with the weighted stochastic approach to index number theory discussed in section "The Stochastic Approach to Index Number Theory" above.

T12: *Quantity reversal test* (quantity weights symmetry test): $P(p^0, p^1, q^0, q^1) = P(p^0, p^1, q^1, q^0)$.

That is, if the quantity vectors for the two periods are interchanged, then the price index remains invariant. This property means that if quantities are used to weight the prices in the index number formula, then the period 0 quantities q^0 and the period 1 quantities q^1 must enter the formula in a symmetric or even-handed manner. Funke and Voeller (1978, p. 3) introduced this test; they called it the *weight property*.

The next test proposed by Diewert (1992a, p. 218) is the analogue to T12 applied to quantity indexes:

T13: *Price reversal test* (price weights symmetry test):

$$\left\{\sum_{n=1}^{N} p_i^1 q_i^1 / \sum_{i=1}^{N} p_j^0 q_i^0\right\} / P\left(p^0, p^1, q^0, q^1\right) = \left\{\sum_{i-1}^{N} p_i^0 q_i^1\right\} / P\left(p^1, p^0, q^0, q^1\right)$$

Thus, if we use (5) to define the quantity index Q in terms of the price index P, then it can be seen that T13 is equivalent to the following property for the associated quantity index Q:

$$Q\left(p^0, p^1, q^0, q^1\right) = Q\left(p^1, p^0, q^0, q^1\right). \tag{25}$$

That is, if the price vectors for the two periods are interchanged, then the quantity index remains invariant. Thus if prices for the same good in the two periods are used to weight quantities in the construction of the quantity index, then property T13 implies that these prices enter the quantity index in a symmetric manner.

The next three tests are mean value tests. The following test was proposed by Eichhorn and Voeller (1976, p. 10):

T14: *Mean value test for prices*:

$$\min_i\left(p_i^1/p_i^0 : i = 1, \ldots, N\right) \leq P\left(p^0, p^1, q^0, q^1\right) \leq \max_i\left(p_i^1/p_i^0 : i = 1, \ldots, N\right).$$

That is, the price index lies between the minimum price ratio and the maximum price ratio. Since the price index is supposed to be some sort of an average of the N price ratios, p_i^1/p_i^0, it seems essential that the price index P satisfy this test.

The next test proposed by Diewert (1992a, p. 219) is the analogue to T14 applied to quantity indexes:

T15: *Mean value test for quantities*:

$$\min_i\left(q_i^1/q_i^0 : i = 1, \ldots, n\right) \leq \left\{V^1/V^0\right\} / P\left(p^0, p^1, q^0, q^1\right) \leq \max_i\left(q_i^1/q_i^0 : i = 1, \ldots, n\right)$$

where V^t is the period t value aggregate $V^t \equiv \sum_{n=1}^{N} p_n^t q_n^t$ for $t = 0, 1$. Using (5) to define the quantity index Q in terms of the price index P, we see that T15 is equivalent to the following property for the associated quantity index Q:

$$\begin{aligned} &\min_i \left(q_i^1/q_i^0 : i = 1, \ldots, N\right) \\ &\quad \leq Q\left(p^0, p^1, q^0, q^1\right) \\ &\quad \leq \max_i \left(q_i^1/q_i^0 : i = 1, \ldots, N\right). \end{aligned} \tag{26}$$

That is, the implicit quantity index Q defined by P lies between the minimum and maximum rates of growth q_i^1/q_i^0 of the individual quantities.

In section "Fixed Basket Approaches," it was argued that it was very reasonable to take an average of the Laspeyres and Paasche price indexes as a single best measure of overall price change. This point of view can be turned into a test:

T16: *Paasche and Laspeyres bounding test*: The price index P lies between the Laspeyres and Paasche indexes, PL and PP, defined by (8) and (9) above.

Bowley (1901, p. 227) and Fisher (1922, p. 403) both endorsed this property for a price index.

Our final four tests are monotonicity tests; that is, how should the price index $P(p^0, p^1, q^0, q^1)$ change as any component of the two price vectors p^0 and p^1 increases or as any component of the two quantity vectors q^0 and q^1 increases.

T17: *Monotonicity in current prices*: $P(p^0, p^1, q^0, q^1) < P(p^0, p^2, q^0, q^1)$ if $p^1 < p^2$.

That is, if some period 1 price increases, then the price index must increase, so that $P(p^0, p^1, q^0, q^1)$ is increasing in the components of p^1. This property was proposed by Eichhorn and Voeller (1976, p. 23) and it is a very reasonable property for a price index to satisfy.

T18: *Monotonicity in base prices*: $P(p^0, p^1, q^0, q^1) > P(p^2, p^1, q^0, q^1)$ if $p^0 < p^2$.

That is, if any period 0 price increases, then the price index must decrease, so that $P(p^0, p^1, q^0, q^1)$ is decreasing in the components of p^0. This very reasonable property was also proposed by Eichhorn and Voeller (1976, p. 23).

T19: *Monotonicity in current quantities*: *if* $q^1 < q^2$, *then*

$$\left\{\sum_{i=1}^{N} p_i^1 q_i^1 / \sum_{i=1}^{N} p_i^0 q_i^0\right\} / P\left(p^0, p^1, q^0, q^1\right) < \left\{\sum_{i=1}^{N} p_i^1 q_i^2 / \sum_{i=1}^{N} p_i^0 q_i^0\right\} / P\left(p^0, p^1, q^0, q^2\right).$$

T20: *Monotonicity in base quantities: if* $q^0 < q^2$, *then*

$$\left\{\sum_{i=1}^{N} p_i^1 q_i^1 / \sum_{i=1}^{N} p_i^0 q_i^0\right\} / P\left(p^0, p^1, q^0, q^1\right) > \left\{\sum_{i=1}^{N} p_i^1 q_i^1 / \sum_{i=1}^{N} p_i^0 q_i^2\right\} / P\left(p^0, p^1, q^2, q^1\right).$$

If we define the implicit quantity index Q that corresponds to P using (1), we find that T19 translates into the following inequality involving Q:

$$Q\left(p^0, p^1, q^0, q^1\right) < Q\left(p^0, p^1, q^0, q^2\right) \text{if } q^1 < q^2. \tag{27}$$

That is, if any period 1 quantity increases, then the implicit quantity index Q that corresponds to the price index P must increase. Similarly, we find that T20 translates into:

$$Q\left(p^0, p^1, q^0, q^1\right) > Q\left(p^0, p^1, q^2, q^1\right) \text{if } q^0 < q^2. \tag{28}$$

That is, if any period 0 quantity increases, then the implicit quantity index Q must decrease. Tests T19 and T20 are due to Vogt (1980, p. 70).

Diewert (1992a, p. 221) showed that the only index number formula $P(p^0, p^1, q^0, q^1)$ which satisfies tests T1–T20 is the Fisher ideal price

index P_F defined earlier by (14), as the geometric mean of the Laspeyres and Paasche price indexes.

P_F satisfies yet another test, T21, which was Fisher's (1921, p. 534, 1922, pp. 72–81) third reversal test (the other two being T9 and T11):

T21: *Factor reversal test* (functional form symmetry test):

$$P(p^0,p^1,q^0,q^1)P(q^0,q^1,p^0,p^1) = \sum_{i=1}^{N} p_i^1 q_i^1 / \sum_{i=1}^{N} p_i^0 q_i^0.$$

A justification for this test is the following one: if $P(p^0, p^1, q^0, q^1)$ is a good functional form for the price index, then if we reverse the roles of prices and quantities, $P(q^0, q^1, p^0, p^1)$ ought to be a good functional form for a quantity index (which seems to be a correct argument) and thus the product of the price index $P(p^0, p^1, q^0, q^1)$ and the quantity index $Q(p^0, p^1, q^0, q^1) = P(q^0, q^1, p^0, p^1)$ ought to equal the value ratio, V^1/V^0. The second part of this argument does not seem to be valid and thus many researchers over the years have objected to the factor reversal test. However, if one is willing to embrace T21 as a basic test, Funke and Voeller (1978, p. 180) showed that the only index number function $P(p^0, p^1, q^0, q^1)$ which satisfies T1 (positivity), T11 (time reversal test), T12 (quantity reversal test) and T21 (factor reversal test) is the Fisher ideal index P_F defined by (14).

Other characterizations of the Fisher price index can be found in Funke and Voeller (1978) and Balk (1985; p. 1995).

The Fisher price index P_F satisfies all 20 of the tests listed above. Which tests do other commonly used price indexes satisfy? Recall the Laspeyres index P_L defined by (8), the Paasche index P_P defined by (9) and the Törnqvist-Theil index P_T defined by (22). Straightforward computations show that the Paasche and Laspeyres price indexes fail only the three reversal tests, T11, T12 and T13. Since the quantity and price reversal tests, T12 and T13, are somewhat controversial and hence can be discounted, the test performance of P_L and P_P seems at first sight to be quite good. However, the failure of the time reversal test, T11, is a severe limitation associated with the use of these indexes.

The Törnqvist-Theil price index P_T fails nine tests: T4 (the fixed basket test), the quantity and price reversal tests T12 and T13, T15 (the mean value test for quantities), T16 (the Paasche and Laspeyres bounding test) and the four monotonicity tests T17–T20. Thus the Törnqvist-Theil index is subject to a rather high failure rate from the perspective of this particular axiomatic approach to index number theory.

However, it could be argued that the list of tests or axioms that was used to establish the superiority of the Fisher ideal index might have been chosen to favour this index. Thus Diewert (2004), following the example of Walsh (1901, pp. 104–05) and Vartia (1976), developed a set of axioms for price indexes of the form $P(p^0, p^1, v^0, v^1)$ where v^0 and v^1 are vectors of expenditures on the N commodities in the index and these vectors replace the quantity vectors q^0 and q^1 as weighting vectors for the prices. In this new axiomatic framework, the Törnqvist-Theil index P_T emerged as the best.

The consistency and independence of various bilateral index number tests was studied in some detail by Eichhorn and Voeller (1976). Our conclusion at this point echoes that of Frisch (1936): the test approach to index number theory, while extremely useful, does not lead to a single unique index number formula. However, two test approaches that take alternative approaches to the methods for weighting prices do lead to the Fisher and Törnqvist-Theil indexes as the best in their respective axiomatic frameworks.

For additional material on the test approach to bilateral index number theory, see Balk (1995), Reinsdorf and Dorfman (1999), Balk and Diewert (2001), Vogt and Barta (1997), and Reinsdorf (2007).

In the following three sections, we consider various economic approaches to index number theory. In the economic approach to price index theory, quantity vectors are no longer regarded as being exogenous variables; rather, they are regarded as solutions to various economic optimization problems.

The Economic Approach to Price Indexes

Before a definition of a microeconomic price index is presented, it is necessary to make a few preliminary definitions.

Let $F(q)$ be a function of N variables, $q \equiv (q_1, \ldots, q_N)$. In the consumer context, F represents a consumer's preferences; i.e. if $F(q^2) > F(q^1)$, then the consumer prefers the commodity vector q^2 over q^1. In this context, F is called a *utility function*. In the producer context, $F(q)$ might represent the output that could be produced using the input vector q. In this context, F is called a *production function*. In order to cover both contexts, we follow the example of Diewert (1976) and call F an *aggregator function*.

Suppose the consumer or producer faces prices $p \equiv (p_1, \ldots, p_N)$ for the N commodities. Then the economic agent will generally find it is useful to minimize the cost of achieving at least a given utility or output level u; we define the *cost function* or *expenditure function* C as the solution to this minimization problem:

$$C(u,p) \equiv \min_q \{p.q : F(q) \geq u\} \tag{29}$$

where $p.q \equiv \sum_{n-1}^{N} p_n q_n$ is the inner product of the price vector p and quantity vector q.

Note that the cost function depends on $1+N$ variables; the utility or output level u and the N commodity prices in the vector p. Moreover, the functional form for the aggregator function F completely determines the functional form for C.

We say that an aggregator function is *neoclassical* if F is: (i) continuous, (ii) positive; i.e. $F(q) > 0$ if $q >> 0_N$ and (iii) linearly homogeneous; that is, $F(\lambda q) = \lambda F(q)$ if $\lambda > 0$. If F is neoclassical, then the corresponding cost function $C(u, p)$ equals u times the unit cost function, $c(p) \equiv C(1, p)$, where $c(p)$ is the minimum cost of producing one unit of utility or output; that is,

$$C(u,p) = uC(1,p) = uc(p). \tag{30}$$

Shephard (1953) formally defined an aggregator function F to be *homothetic* if there exists an increasing continuous function of one variable g such that $g[F(q)]$ is neoclassical. However, the concept of homotheticity was well known to Frisch (1936) who termed it expenditure proportionality. If F is homothetic, then its cost function C has the following decomposition:

$$\begin{aligned} C(u,p) &\equiv \min_q \{p \cdot q : F(q) \geq u\} \\ &= \min_q \{p \cdot q : g[F(q)] \geq g(u)\} \\ &= g(u)c(p) \end{aligned} \tag{31}$$

where $c(p)$ is the unit cost function that corresponds to $g[F(q)]$.

Let $p^0 >> 0_N$ and $p^1 >> 0_N$ be positive price vectors pertaining to periods or observations 0 and 1. Let $q > 0_N$ be a non-negative, non-zero reference quantity vector. Then the Konüs (1924) *price index* or *cost of living index* is defined as:

$$P_K(p^0, p^1, q) \equiv C[F(q), p^1] / C[F(q), p^0]. \tag{32}$$

In the consumer (producer) context, P_K may be interpreted as follows. Pick a reference utility (output) level $u \equiv F(q)$. Then $P_K(p^0, p^1, q)$ is the minimum cost of achieving the utility (output) level u when the economic agent faces prices p^1 relative to the minimum cost of achieving the same u when the agent faces prices p^0. If $N = 1$ so that there is only one consumer good (or input), then it is easy to show that $P_K(p_1^0, p_1^1, q_1) = p_1^1 q_1 / p_1^0 q_1 = p_1^1 / p_1^0$.

Using the fact that a cost function is linearly homogeneous in its price arguments, it can be shown that P_K has the following homogeneity property: $P_K(p^0, \lambda p^1, q) = \lambda P_K(p^0, p^1, q)$ for $\lambda > 0$ which is analogous to the proportionality test T5 in the previous section. P_K also satisfies $P_K(p^1, p^0, q) = 1/P_K(p^0, p^1, q)$ which is analogous to the time reversal test, T11.

Note that the functional form for P_K is completely determined by the functional form for the aggregator function F, which determines the functional form for the cost function C.

In general, P_K depends not only on the two price vectors p^0 and p^1, but also on the reference vector q. Malmquist (1953), Pollak (1983), and Samuelson and Swamy (1974) have shown that P_K is independent of q and is equal to a ratio of unit cost functions, $c(p^1)/c(p^0)$, if and only if the aggregator function F is homothetic.

If we knew the consumer's preferences or the producer's technology, then we would know F and we could construct the cost function C and the Konüs price index P_k. However, we generally do not know F or C and thus it is useful to develop *bounds* that depend on observable price and quantity data but do not depend on the specific functional form for F or C.

Samuelson (1947) and Pollak (1983) established the following bounds on P_K.

Let $p^0 >> 0_N$ and $p^1 >> 0_N$. Then for every reference quantity vector $q > 0_N$, we have

$$\min_n\{p_n^1/p_n^0\} \leq P_K(p^0,p^1,q) \leq \max_n\{p_n^1/p_n^0\}; \quad (33)$$

that is, P_K lies between the smallest and largest price ratios. Unfortunately, these bounds are usually too wide to be of much practical use.

To obtain closer bounds, we now assume that the observed quantity vectors for the two periods, $q^i \equiv (q_1^i,\ldots,q_N^i), i = 0, 1,$ are solutions to the producer's or consumer's cost minimization problems; that is, we assume:

$$p^i \cdot q^i = C[F(q^i),p^i], p^i >> 0_N, q^i > 0_N, i = 0, 1. \quad (34)$$

Given the above assumptions, we now have two natural choices for the reference quantity vector q that occurs in the definition of $P_K(p^0, p^1, q)$: q^0 or q^1. The *Laspeyres–Konüs price index* is defined as $P_K(p^0, p^1, q^0)$ and the *Paasche–Konüs price index* is defined as $P_K(p^0, p^1, q^1)$.

Under the assumption of cost minimizing behaviour (34), Konüs (1924) established the following bounds:

$$P_K(p^0,p^1,q^0) \leq p^1 \cdot q^0/p^0 \cdot q^0 \equiv P_L(p^0,p^1,q^0,q^1); \quad (35)$$

$$P_K(p^0,p^1,q^1) \geq p^1 \cdot q^1/p^0 \cdot q^1 \equiv P_P(p^0,p^1,q^0,q^1), \quad (36)$$

where P_L and P_P are the Laspeyres and Paasche price indexes defined earlier by (8) and (9). If in addition, the aggregator function is homothetic, then Frisch (1936) showed that for any reference vector $q > 0_N$,

$$P_P \equiv p^1 \cdot q^1/p^0 \cdot q^1 \leq P_K(p^0,p^1,q) \leq p^1 \cdot q^0/p^0 \cdot q^0 \equiv P_L. \quad (37)$$

In the consumer context, it is unlikely that preferences will be homothetic; hence the bounds (37) cannot be justified in general. However, Konüs (1924) showed that bounds similar to (37) would hold even in the general non-homothetic case, provided that we choose a reference vector $q \equiv \lambda q^0 + (1 - \lambda)q^1$ which is a λ, $(1 - \lambda)$ weighted average of the two observed quantity points. Specifically, Konüs showed that there exists a λ between 0 and 1 such that if $P_P \leq P_L$, then

$$P_P \leq P_K[p^0,p^1,\lambda q^0 + (1-\lambda)q^1] \leq P_L \quad (38)$$

or if $P_p > P_L$, then

$$P_L \leq P_K[p^0,p^1,\lambda q^0 + (1-\lambda)q^1] \leq P_P. \quad (39)$$

The bounds on the microeconomic price index P_K given by (37) in the homothetic case and (38)–(39) in the non-homothetic case are the best bounds that we can obtain without making further assumptions on F. In the time series context, the bounds given by (38) or (39) are usually quite satisfactory: the Paasche and Laspeyres price indexes for consecutive time periods will usually differ by less than one per cent (and hence taking the Fisher geometric average will generally suffice for most practical purposes). However, in the cross-section context where the observations represent, for example, production data for two producers in the same industry but in different regions, the bounds are often not very useful since P_L and P_P can differ by 50 per cent or more in the cross-sectional context: see Ruggles (1967) and Hill (2006a).

For generalizations of the above single household theory to many households, see Pollak (1980, p. 276, 1981, p. 328), Diewert (1983a, 2001) and in ILO (2004, ch. 18).

In section "Exact and Superlative Indexes," we will make additional assumptions on the aggregator function F or its cost function dual C that will enable us to determine P_K exactly. Before we do this, in the next section we will define various quantity indexes that have their origins in microeconomic theory.

Economic Approaches to Quantity Indexes

In the one commodity case, a natural definition for a quantity index is q_1^1/q_1^0, the ratio of the single quantity in period 1 to the corresponding quantity in period 0. This ratio is also equal to the expenditure ratio, $p_1^1 q_1^1/p_1^0 q_1^0$, divided by the price ratio, $p_1^1 p_1^0$. This suggests that in the N commodity case a reasonable definition for a quantity index would be the expenditure ratio divided by the Konüs price index, P_K. This type of index was suggested by Pollak (1983). Thus the *Konüs-Pollak quantity index*, Q_K, is defined by:

$$\begin{aligned} &Q_K(p^0,p^1,q^0,q^1,q) \\ &\quad \equiv p^1.q^1/p^0.q^0 P_K(p^0,p^1,q) \\ &\quad = \{C[F(q^1),p^1]/C[F(q),p^1]\}/ \\ &\quad \{C[C(q^0),p^0]/C[F(q),p^0]\} \end{aligned} \tag{40}$$

where the second line follows from the definition of P_K, (32), and the assumption of cost minimizing behaviour in the two periods, (34).

The definition of Q_K depends on the reference vector q which appears in the definition of P_K. The general definition of Q_K simplifies considerably if we choose the reference q to be q^0 or q^1. Thus define the *Laspeyres-Konüs* quantity index as

$$\begin{aligned} &Q_K(p^0,p^1,q^0,q^1,q^0) \\ &\quad \equiv C[F(q^1),p^1]/C[F(q^0),p^1] \end{aligned} \tag{41}$$

and the *Paasche-Konüs* quantity index as

$$\begin{aligned} &Q_K(p^0,p^1,q^0,q^1,q^0) \\ &\quad \equiv C[F(q^1),p^0]/C[F(q^0),p^0]. \end{aligned} \tag{42}$$

The indexes defined by (41) and (42) are special cases of another class of quantity indexes. For any reference price vector $p >> 0_N$, define the *Allen* (1949) *quantity index* by

$$Q_A(q^0,q^1,p) \equiv C[F(q^1),p]/C[F(q^0),p]. \tag{43}$$

If p is chosen to be p^0, (43) becomes (42) and if $p = p^1$, then (43) becomes (41). Using the properties of cost functions, it can be shown that if $F(q^1) \geq F(q^0)$, then $Q_A(q^0, q^1, p) \geq 1$, while if $F(q^1) \leq F(q^0)$, then $Q_A(q^0, q^1, p) \leq 1$. Thus the Allen quantity index correctly indicates whether the commodity vector q^1 is larger or smaller than q^0. It can also be seen that Q_A satisfies a counterpart to the time reversal test; that is, $Q_A(q^1, q^0, p) = 1/Q_A(q^0, q^1, p)$.

Just as the price index P_K depended on the unobservable aggregator function, so also do the quantity indexes Q_K and Q_A. Thus it is useful to develop bounds for the quantity indexes that do not depend on the particular functional form for F.

Samuelson (1947) and Allen (1949) established the following bounds for (41) and (42):

$$\begin{aligned} Q_A(q^0,q^1,p^0) &= Q_K(p^0,p^1,q^0,q^1,q^1) \\ &\leq p^0 \cdot q^1/p^0 \cdot q^0 \equiv Q_L; \end{aligned} \tag{44}$$

$$\begin{aligned} Q_A(q^0,q^1,p^0) &= Q_K(p^0,p^1,q^0,q^1,q^0) \\ &\leq p^1.q^1/p^1.q^0 \equiv Q_P. \end{aligned} \tag{45}$$

Note that the observable *Laspeyres* and *Paasche quantity indexes*, Q_L and Q_P, appear on the right hand sides of (44) and (45).

Diewert (1981), utilizing some results of Pollak (1983) and Samuelson and Swamy (1974), established the following results: if the

underlying aggregator function F is neoclassical and (32) holds, then for all $P >> 0_N$ and $q >> 0_N$,

$$\begin{aligned} Q_P &\leq Q_A(q^0, q^1, p) \\ &= Q_K(p^0, p^1, q^0, q^1, q) \\ &\leq F(q^1)/F(q^0) \leq Q_L \end{aligned} \tag{46}$$

Thus if the aggregator function F is neoclassical, then the Allen quantity index for all reference vectors p equals the Konüs quantity index for all reference quantity vectors q, which in turn equals the ratio of aggregates, $F(q^1)/F(q^0)$. Moreover, Q_A and Q_K are bounded from below by the Paasche quantity index Q_P, and bounded from above by the Laspeyres quantity index Q_L in the neoclassical case.

In the general non-homothetic case, Diewert (1981) showed that there exists a λ between 0 and 1 such that $Q_K(p^0, p^1, q^0, q^1, \lambda q^0 + (1 - \lambda)q^1]$ lies between QP and QL and there exists a λ^* between 0 and 1 such that $Q_A(q^0, q^1, \lambda^* p^0 + (1 - \lambda^*)p^1)$ also lies between Q_P and Q_L. Thus the observable Paasche and Laspeyres quantity indexes bound both the Konüs quantity index and the Allen quantity index, provided that we choose appropriate reference vectors between q^0 and q^1 and p^0 and p^1 respectively.

Using the linear homogeneity property of the cost function in its price arguments, we can show that the Konüs price index has the desirable homogeneity property, $P_K(p^0, \lambda p^0, q) = \lambda$ for all $\lambda > 0$; that is, if period 1 prices are proportional to period 0 prices, then P_K equals this common proportionality factor. It would be desirable for an analogous homogeneity property to hold for quantity indexes.

Unfortunately, it is not in general true that $Q_K(q^0, \lambda q^0, p^0, p^1, q) = \lambda$ or that $Q_A(q^0, \lambda q^0, p) = \lambda$. Thus we turn to a third economic approach to defining a quantity index which has the desirable quantity proportionality property.

Let q^1 and q^2 be the observable quantity vectors in the two situations as usual, let $F(q)$ be an increasing, continuous aggregator function, and let $q >> 0$ be a reference quantity vector. Then the *Malmquist* (1953) *quantity index QM* is defined as:

$$Q_M(q^0, q^1, q) \equiv D[F(q), q^1]/D[F(q), q^0] \tag{47}$$

where $D(u, q^t) \equiv \max_k\{k : F(q^t/k) \geq u, \quad k > 0\}$ is the *deflation* or *distance function* which corresponds to F. Thus $D[F(q), q^1]$ is the biggest number which will just deflate the quantity vector q^1 onto the boundary of the utility (or production) possibilities set $\{z: F(z) \geq F(q)\}$ indexed by the reference quantity vector q while $D[F(q), q^0]$ is the biggest number which will just deflate the quantity vector q^0 onto the set $\{z: F(z) \geq F(q)\}$ and QM is the ratio of these two deflation factors. Note that there is no optimization problem involving prices in the definition of the Malmquist quantity index, but the definition of the distance function involves certain deflation problems that can be interpreted as technical efficiency optimization problems.

Q_M depends on the unobservable aggregator function F and as usual, we are interested in bounds for Q_M.

Diewert (1981) showed that Q_M satisfied bounds analogous to (33); that is,

$$\begin{aligned} \min_n\{q_n^1/q_n^0\} &\leq Q_M(q^0, q^1, q) \\ &\leq \max_n\{q_n^1/q_n^0\} \end{aligned} \tag{48}$$

As noted above, the assumption of cost minimizing behaviour is not required in order to define the Malmquist quantity index or to establish the bounds (46).

However, in order to establish the following bounds due to Malmquist (1953) for Q_M, we do need the assumption of cost-minimizing behaviour (32) for the two periods under consideration, and we require the reference vector q to be q^0 or q^1:

$$Q_M(q^0, q^1, q^0) \leq p^0 \cdot q^1 / p^0 \cdot q^0 \equiv Q_L; \tag{49}$$

$$Q_M(q^0, q^1, q^1) \leq p^1 \cdot q^1 / p^1 \cdot q^0 \equiv Q_P. \tag{50}$$

Diewert (1981) showed that, under the hypothesis of cost-minimizing behaviour, there exists a λ between 0 and 1 such that $Q_M(q^0, q^1, \lambda q^0 + (1 - \lambda)q^1)$ lies between Q_P and Q_L. Thus the Paasche and Laspeyres quantity indexes provide bounds for a Malmquist quantity index for some reference

indifference or product surface indexed by a quantity vector which is a λ, $(1-\lambda)$ weighted average of the two observable quantity vectors, q^0 and q^1.

Pollak (1983) showed that, if F is neoclassical, then we can extend the string of equalities in (46) to include the Malmquist quantity index $Q_M(q^0, q^1, q)$, for any reference quantity vector q. Thus, in the case of a linearly homogeneous aggregator function, all three theoretical quantity indexes coincide and this common theoretical index is bounded from below by the Paasche quantity index Q_P and bounded from above by the Laspeyres quantity index Q_L.

In the general case of a non-homothetic aggregator function, our best theoretical quantity index, the Malmquist index, is also bounded by the Paasche and Laspeyres indexes, provided that we choose a suitable reference quantity vector. In order to improve upon the bounding approach, Caves et al. (1982b) show that, if one is willing to assume optimizing behaviour and make certain functional form assumptions about the underlying technology, then it is possible to obtain exact expressions for the Malmquist quantity index.

We noted in the price index context that the Paasche and Laspeyres price indexes were usually quite close in the time series context. A similar remark also applies to the Paasche and Laspeyres quantity indexes. Thus taking an average of the Paasche and Laspeyres indexes, such as the Fisher price and quantity indexes, will generally approximate underlying microeconomic price and quantity indexes sufficiently accurately for most practical purposes. However, this observation does not apply to the cross-sectional context, where the Paasche and Laspeyres indexes can differ widely. In the following section, we offer another microeconomic justification for using the Fisher indexes that also applies in the context of making interregional and cross-country comparisons.

Exact and Superlative Indexes

Assume that the producer or consumer is maximizing a neoclassical aggregator function f subject to a budget constraint during the two periods. Under these conditions, it can be shown that the economic agent is also minimizing cost subject to a utility or output constraint. Moreover, the cost function C that corresponds to f can be written as $C[f(q), p] = f(q)c(p)$ where c is the unit cost function (see (28) above).

Suppose a bilateral price index $P(p^0, p^1, q^0, q^1)$ and the corresponding quantity index $Q(p^0, p^1, q^0, q^1)$ that satisfy (5) are given. The quantity index Q is defined to be *exact* for a neoclassical aggregator function f with unit cost dual c if for every $P^0 >> 0_N$, $P^1 >> 0_N$ and $q^i >> 0_N$ which is a solution to the aggregator maximization problem $\max_q\{f(q) : p^i.\ q \leq p^i.\ q^i\} = f(q^i) > 0$ for $i = 0, 1$, we have

$$Q(p^0, p^1, q^0, q^1) = f(q^1)/f(q^0). \tag{51}$$

Under the same hypothesis, the price index P is *exact* for f and c if we have

$$P(p^0, p^1, q^0, q^1) = c(p^1)/c(p^0). \tag{52}$$

In (51) and (52), the price and quantity vectors are not regarded as being independent. The *pi* can be independent, but the *qi* are solutions to the corresponding aggregator maximization problem involving p^i, for $i = 0,1$. Note that, if Q is exact for a neoclassical f, then Q can be interpreted as a Konüs, Allen or Malmquist quantity index and the corresponding P defined implicitly by (5) can be interpreted as a Konüs price index.

The concept of exactness is due to Konüs and Byushgens (1926). Below, we shall give some examples of exact index number formulae. Additional examples may be found in Afriat (1972), Pollak (1983), Samuelson and Swamy (1974), and Diewert (1976, 1992b).

Konüs and Byushgens (1926) showed that Irving Fisher's ideal price index P_F defined by (14) and the corresponding quantity index Q_F defined implicitly by (5) are exact for the homogeneous quadratic aggregator function f defined by

$$f(q_1, \ldots, q_N) \equiv \left(\sum_{n-1}^{N}\sum_{m-1}^{N} a_{nm} q_n q_m\right)^{1/2} \equiv (q\ .\ Aq)^{1/2} \tag{53}$$

where $A \equiv [a_{nm}]$ is a symmetric $N \times N$ matrix of constants. Thus, under the assumption of maximizing behaviour, we can show that $f(q^1)/f(q^0) = Q_F$ and $c(p^1)/c(p^0) = P_F$ where f is defined by (51) and c is the unit cost function that corresponds to f. The important point to note is that f depends on $N(N+1)/2$ unknown a_{nm} parameters but we do not need to know these parameters in order to be able to calculate $f(q^1)/f(q^0)$and $c(p^1)/c(p^0)$.

Diewert (1976) showed that the Törnqvist–Theil price index P_T defined by (22) is exact for the unit cost function $c(p)$ defined by:

$$\ln c(p) \equiv \alpha_0 + \sum_{n=1}^{N} \alpha_n \ln p_n + (1/2) \sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_{mn} \ln p_m \ln p_n \tag{54}$$

where the parameters α_n and α_{mn} satisfy the following restrictions:

$$\sum_{n=1}^{N} \alpha_n = 1, \sum_{n=1}^{N} \alpha_{mn} = 0 \text{ for } m = 1, \ldots, N \text{ and } \alpha_{mn} = \alpha_{mn} \text{ for all } m, n. \tag{55}$$

Thus we may calculate $c(p^1)/c(p^0) = P_T$ and $f(q^1)/f(q^0) = p^1 \cdot q^1 = p^0 \cdot q^0 P_T \equiv Q_T$ where c is the unit cost function defined by (54), f is the aggregator function which corresponds to this c, and QT is the implicit Törnqvist-Theil quantity index. Note that we do not have to know the parameters αn and αmn in order to evaluate $c(p^1)/c(p^0)$ and $f(q^1)/f(q^0)$.

The unit cost function defined by (54) is the *translog* unit cost function defined by Christensen et al. (1971). Since PT is exact for this translog functional form, P_T is sometimes called the *translog price index*.

Define the following family of quantity indexes Qr that depend on a number, $r \neq 0$:

$$Q_r(p^0, p^1, q^0, q^1) \equiv \left[\sum_{n=1}^{N} s_n^0 (q_n^1/q_n^0)^{r/2}\right]^{1/r} \left[\sum_{m=1}^{N} s_m^1 (q_m^1/q_m^0)^{-r/2}\right]^{-1/r} \tag{56}$$

where $s_n^i \equiv p_n^i q_n^i / p^i \cdot q^i$ is the period i expenditure share for good n. For each $r \neq 0$, define the corresponding implicit price index by:

$$P_r^*(p^0, p^1, q^0, q^1) \equiv p^1 \cdot q^1 / p^0 \cdot q^0 \, Q_r(p^0, p^1, q^0, q^1). \tag{57}$$

A quick algebraic calculation will show that when $r = 2$, $P_2^* = P_F$, the Fisher price index defined by (14) and when r equals 1, P_1^*equals:

$$P_1^* = \sum_{n=1}^{N} p_n^1 (q_n^0 q_n^1)^{1/2} / \sum_{m=1}^{N} p_m^0 (q_m^0 q_m^1)^{1/2} = P_W \tag{58}$$

Where P_W is the Walsh price index defined earlier by (17).

Diewert (1976) showed that Q_r and P_r^* are exact for the *quadratic mean* of *order r aggregator function fr* defined as follows:

$$f_r(q_1, \ldots, q_N) \equiv \left(\sum_{m=1}^{N} \sum_{n=1}^{N} a_{mn} q^{r/2} q^{r/2}\right)^{1/r} \tag{59}$$

Where $A = [amn]$ is a symmetric matrix of constants. Thus the Walsh and Fisher price indexes, PW and PF, are exact for $f1(q)$ and $f2(q)$ respectively, defined by (59) when $r = 1$ and 2.

Diewert (1974) defined a linearly homogeneous function f of N variables to be *flexible* if it could provide a second-order approximation to an arbitrary twice continuously differentiable linearly homogeneous function. It can be shown that f defined by (53), c defined by (54) and (55) and fr

defined by (59) for each $r \neq 0$ are all examples of flexible functional forms.

Let the price and quantity indexes P and Q satisfy the product test equality, (5). Then Diewert (1976) defined P and Q to be *superlative indexes* if either P is exact for a flexible unit cost function c or Q is exact for a flexible aggregator function f.

Thus P_F, P_W, P_T and P_r^* are all superlative price indexes. Thus from the viewpoint of the economic approach to index number theory, all of these indexes can be judged to be equally good.

At this point, it is useful to review the various approaches to bilateral index number theory discussed in the previous sections. In section "Fixed Basket Approaches," it was found that the best average basket approaches led to the Fisher or Walsh price indexes. In section "The Stochastic Approach to Index Number Theory," the index from the viewpoint of the stochastic approach was the Törnqvist-Theil index. In section "The Test Approach to Index Number Theory," the test approach led to the Fisher or the Törnqvist-Theil indexes as being best. Finally, in this section, the economic approach led to the Fisher, Walsh and Fisher or the Törnqvist-Theil indexes as being equally good. *Thus all four major approaches to index number theory led to the same three indexes as being best*. But which one of these three formulae, P_F, P_W and P_T, should we choose? Fortunately, it does not matter very much which of these formulae we choose to use in applications; they will all give the same answer to a reasonably high degree of approximation. Diewert (1978, p. 889) showed that all known superlative index number formulae approximate each other to the second order when each index is evaluated at an equal price and quantity point. This means the P_F, P_W, P_T and each P_r^* have the same first and second order partial derivatives with respect to all 4 N arguments when the derivatives are evaluated at a point where $p^0 = p^1$ and $q^0 = q^1$. A similar string of equalities also holds for the corresponding implicit quantity indexes defined using the product test (5). In fact, these derivative equalities are still true provided that $p^1 = \lambda p^0$ and $q^1 = \mu q^0$ for any numbers $\lambda > 0$ and $\mu > 0$. However, although Diewert's approximation result is mathematically true, Hill (2006) has shown that superlative indexes of the form P_r^* for r very large in magnitude do not necessarily empirically approximate the standard superlative indexes P_F, P_W and P_T very closely. But these standard superlative indexes typically approximate each other to something less than 0.2 per cent in the time series context and to about two per cent in the cross-section context; see Fisher (1922), Ruggles (1967), Diewert (1978, pp. 894–5) and Hill (2006) for empirical evidence on this point.

Diewert (1978) also showed that the Paasche and Laspeyres indexes approximate the superlative indexes to the first order at an equal price and quantity point. In the time series context, for adjacent periods, the Paasche and Laspeyres price indexes typically differ by less than 0.5 per cent; hence these indexes may provide acceptable approximations to a superlative index.

After consideration of the case of two observations at length, the many- observation case is considered in the following two sections.

The Fixed Base Versus the Chain Principle

In this section, the merits of using the chain system for constructing price indexes in the time series context versus using the fixed base system are discussed.

The chain system, introduced independently into the economics literature by Lehr (1885, pp. 45–6) and Marshall (1887, p. 373), measures the change in prices going from one period to a subsequent period using a bilateral index number formula involving the prices and quantities pertaining to the two adjacent periods. These one period rates of change (the links in the chain) are then cumulated to yield the relative levels of prices over the entire period under consideration. Thus, if the bilateral price index is P, the chain

system generates the following pattern of price levels for the first three periods:

$$1, P(p^0, p^1, q^0, q^1), \quad P(p^0, p^1, q^0, q^1) \quad P(p^1, p^2, q^1, q^2). \tag{60}$$

On the other hand, the fixed base system of price levels using the same bilateral index number formula P simply computes the level of prices in period t relative to the base period 0 as $P(p^0, p^t, q^0, q^t)$. Thus the fixed base pattern of price levels for periods 0,1 and 2 is:

$$1, P(p^0, p^1, q^0, q^1) \quad P(p^0, p^2, q^0, q^2). \tag{61}$$

Due to the difficulties involved in obtaining current period information on quantities (or equivalently, on expenditures), as was indicated in section "Fixed Basket Approaches," many statistical agencies loosely base their consumer price index on the use of the Laspeyres formula and the fixed base system. Therefore, it is of some interest to look at some of the possible problems associated with the use of fixed base Laspeyres indexes.

The main problem with the use of the fixed base Laspeyres index is that the period 0 fixed basket of commodities that is being priced out in period t can often be quite different from the period t basket. Thus, if there are systematic *trends* in at least some of the prices and quantities in the index basket, the fixed base Laspeyres price index $P_L(p^0, p^t, q^0, q^t)$ can be quite different from the corresponding fixed base Paasche price index, $P_P(p^0, p^t, q^0, q^t)$. This means that both indexes are likely to be an inadequate representation of the movement in average prices over the time period under consideration.

As Hill (1988) noted, the fixed base Laspeyres quantity index cannot be usedfor ever: eventually, the base period quantities q^0 are so far removed from the current period quantities q^t that the base must be changed. Chaining is merely the limiting case where the base is changed each period.

The main advantage of the chain system is that under normal conditions, chaining will reduce the spread between the Paasche and Laspeyres indexes; see Diewert (1978, p. 895) and Hill (1988, 1993, pp. 387–8). These two indexes each provide an asymmetric perspective on the amount of price change that has occurred between the two periods under consideration, and it could be expected that a single point estimate of the aggregate price change should lie between these two estimates. Thus the use of either a chained Paasche or Laspeyres index will usually lead to a smaller difference between the two and hence to estimates that are closer to the 'truth'.

Hill (1993, p. 388), drawing on the earlier research of Szulc (1983) and Hill (1988, pp. 136–7), noted that it is not appropriate to use the chain system when prices oscillate or 'bounce', to use Szulc's (1983, p. 548) term. This phenomenon can occur in the context of regular seasonal fluctuations or in the context of price wars. However, in the context of roughly monotonically changing prices and quantities, Hill (1993, p. 389) recommended the use of chained symmetrically weighted indexes. The Fisher, Walsh and Törnqvist-Theil indexes are examples of symmetrically weighted indexes.

It is possible to be more precise about the conditions under which one should chain or not chain. Following arguments due to Walsh (1901, p. 206, 1921a, pp. 84–5) and Fisher (1911, pp. 204 and 423–4), one should chain if the prices and quantities pertaining to adjacent periods are *more similar* than the prices and quantities of more distant periods, since this strategy will lead to a narrowing of the spread between the Paasche and Laspeyres indexes at each link. Of course, one needs a measure of how similar the prices and quantities pertaining to two periods are. The similarity measures could be *relative* ones or *absolute* ones. In the case of absolute comparisons, two vectors of the same dimension are similar if they are identical and dissimilar otherwise. In the case of relative comparisons, two vectors are similar if they are proportional and dissimilar if they are non-proportional. Once a similarity measure has been defined, the prices and quantities of each period can be compared with each other using this measure, and a 'tree' or path that links all the observations can be constructed where the most similar observations are compared with each other using a bilateral index number formula.

Fisher (1922, pp. 271–6) informally suggested this strategy. However, the more recent literature on this approach is due to Robert Hill. Initially, Hill (1999a, b, 2001) defined the price structures between the two countries to be more dissimilar the bigger is the spread between P_L and P_P, that is, the bigger is $\max\{P_L/P_P, P_P/P_L\}$. The problem with this measure of dissimilarity in the price structures of the two countries is that it could be the case that $P_L = P_P$ (so that the Hill measure would register a maximal degree of similarity) but p^0 could be very different from p^t. Thus there is a need for a more systematic study of similarity (or dissimilarity) measures in order to pick the best one that could be used as an input into Hill's (1999a, b, 2001, 2004, 2006b, 2007) spanning tree algorithm for linking observations, see Diewert (2007a).

The method of linking observations explained in the previous paragraph based on the similarity of the price and quantity structures of any two observations may not be practical in a statistical agency context since the addition of a new period may lead to a reordering of the previous links. However, the above 'scientific' method for linking observations may be useful in deciding whether chaining is preferable or whether fixed base indexes should be used while making month-to-month comparisons within a year.

Some index number theorists have objected to the chain principle on the grounds that it has no counterpart in the spatial context:

> They [chain indexes] only apply to intertemporal comparisons, and in contrast to direct indices they are not applicable to cases in which no natural order or sequence exists. Thus the idea of a chain index for example has no counterpart in interregional or international price comparisons, because countries cannot be sequenced in a 'logical' or 'natural' way (there is no $k+1$ nor $k-1$ country to be compared with country k). (von der Lippe 2001, p. 12)

This is of course correct but the approach of Robert Hill leads to a 'natural' set of spatial links. Applying the same approach to the time series context will lead to a set of links between periods which may not be month-to-month but it will in many cases justify year-over-year linking of the data pertaining to the same month.

It is of some interest to determine if there are index number formulae that give the same answer when either the fixed base or chain system is used. If we compare the sequence of chain indexes defined by (60) above with the corresponding fixed base indexes defined by (61), it can be seen that we will obtain the same answer in all three periods if the index number formula P satisfies the following functional equation for all price and quantity vectors:

$$P(p^0,p^2,q^0,q^2) = P(p^0,p^1,q^0,q^1)P(p^1,p^2,q^1,q^2). \tag{62}$$

If a bilateral index number formula P satisfies (62), then P satisfies the *circularity test*, see Westergaard (1890, pp. 218–19) and Fisher (1922, p. 413).

If it is assumed that the index number formula P satisfies certain properties or tests in addition to the circularity test above, then Funke et al. (1979) showed that P must have the following functional form due originally to Konüs and Byushgens (1926, pp. 163–6):

$$\ln P_{KB}(p^0,p^1,q^0,q^1) \equiv \sum_{i=1}^{N} \alpha_i \ln (p_i^1/p_i^0) \tag{63}$$

where the N constants α_i satisfy the following restrictions:

$$\sum_{i=1}^{N} \alpha_i = 1 \text{ and } \alpha_i > 0 \text{ for } i = 1,\ldots,N. \tag{64}$$

Thus, under very weak regularity conditions, the only price index satisfying the circularity test is a weighted geometric average of all the individual price ratios, the weights being constant through time. This result vindicates Irving Fisher's (1922, p. 274) intuition when he asserted that 'the only formulae which conform perfectly to the circular test are index numbers which have *constant weights*.. .'.

The problem with the indexes defined by Konüs and Byushgens is that the individual price

ratios, p_n^1/p_n^0 have weights that are *independent* of the economic importance of commodity n in the two periods under consideration. Put another way, these price weights are independent of the quantities of commodity n consumed or the expenditures on commodity n during the two periods. Hence, these indexes are not really suitable for use by statistical agencies at higher levels of aggregation when expenditure share information is available.

The above results indicate that it is not useful to ask that the price index P satisfy the circularity test *exactly*. However, it is of some interest to find index number formulae that satisfy the circularity test to some degree of *approximation* since the use of such an index number formula will lead to measures of aggregate price change that are more or less the same no matter whether we use the chain or fixed base systems. Irving Fisher (1922, p. 284) found that deviations from circularity using his data-set and the Fisher ideal price index P_F were quite small. This relatively high degree of correspondence between fixed base and chain indexes has been found to hold for other symmetrically weighted formulae like the Walsh index P_W defined earlier. It is possible to give a theoretical explanation for the approximate satisfaction of the circularity test in the time series context for symmetrically weighted index number formulae, such as P_F and P_W. Another symmetrically weighted formula is the Törnqvist-Theil index P_T. Alterman et al. (1999, p. 61) showed that if the logarithmic price ratios $\ln\left(p_n^t/p_n^{t-1}\right)$ trend linearly with time t and the expenditure shares s_n^t also trend linearly with time, then the Törnqvist index P_T will satisfy the circularity test exactly. Since many economic time series on prices and quantities satisfy these assumptions approximately, then the Törnqvist index P_T will satisfy the circularity test approximately. As was noted earlier, the Törnqvist index generally closely approximates the symmetrically weighted Fisher and Walsh indexes, so that for many economic time series (with smooth trends) all three of these symmetrically weighted indexes will satisfy the circularity test to a high enough degree of approximation so that it will not matter whether we use the fixed base or chain principle.

Walsh (1901, p. 401, 1921a, p. 98, 1921b, p. 540) introduced the following useful variant of the circularity test:

$$1 = P\left(p^0,p^1,q^0,q^1\right)\ P\left(p^1,p^2,q^1,q^2\right)\ldots P\left(p^{T-1},p^T,q^{T-1},q^T\right)\ P\left(p^T,p^0,q^T,q^0\right). \quad (65)$$

The motivation for this test is the following. Use the bilateral index formula $P(p^0, p^1, q^0, q^1)$ to calculate the change in prices going from period 0 to 1, use the same formula evaluated at the data corresponding to periods 1 and 2, $P(p^1, p^2, q^1, q^2)$, to calculate the change in prices going from period 1 to 2, …, use $P(p^{T-1}, p^T, q^{T-1}, q^T)$ to calculate the change in prices going from period $T-1$ to T, introduce an artificial period $T+1$ that has exactly the price and quantity of the initial period 0 and use $P(p^0, p^1, q^0, q^1)$ to calculate the change in prices going from period T to 0. Finally, multiply all these indexes together, and since we end up where we started the product of all of these indexes should ideally be 1. Diewert (1993a, p. 40) called this test a *multiperiod identity test*. Note that, if $T = 2$ (so that the number of periods is 3 in total), then Walsh's test reduces to Fisher's (1921, p. 534, 1922, p. 64) *time reversal test*.

Walsh (1901, pp. 423–33) showed how his circularity test could be used in order to evaluate how 'good' any bilateral index number formula was. What he did was invent artificial price and quantity data for five periods, and he added a sixth period that had the data of the first period. He then evaluated the right-hand side of (65) for various bilateral formula, $P(p^0, p^1, q^0, q^1)$, and determined how far from unity the results were. His best formulae had products that were close to 1. Fisher (1922, p. 284) later used this methodology as well.

This same framework is often used to evaluate the efficacy of chained indexes versus their direct counterparts. Thus if the right hand side of (65) turns out to be different from unity, the chained indexes are said to suffer from 'chain drift'. If a formula suffers from chain drift, it is sometimes recommended that fixed base indexes be used in place of chained ones. However, this advice, if accepted, would *always* lead to the adoption of fixed base indexes, provided that the bilateral

index formula satisfies the identity test, $P(p^0, p^0, q^0, q^0) = 1$. Thus it is not recommended that Walsh's circularity test be used to decide whether fixed base or chained indexes should be calculated. However, it is fair to use Walsh's circularity test as he originally used it, namely, as an approximate method for deciding how good a particular index number formula is. In order to decide whether to chain or use fixed base indexes, one should decide on the basis of how similar the observations being compared are, and choose the method which will best link up the most similar observations.

Robert Hill's method for linking observations can be regarded as a multilateral index number method, one which is based on a suitable bilateral formula, a measure of the similarity of any two price and quantity vectors and an algorithm for linking the observations via a path that links the most similar observations. In the following section, we review some other multilateral methods.

Multilateral Indexes

Assume that there are I positive price vectors $p^i \equiv (p_1^i, \ldots, p_N^i)$ and I quantity vectors $q^i \equiv (q_1^i, \ldots, q_N^i)$ with $p^i\, q^i > 0$ for $i = 1,\ldots,I$. We wish to find $2I$ positive numbers P^i (price indexes) and Q^i (quantity indexes) such that $PiQi = pi \cdot qi$ for $i = 1, \ldots,I$. The I data points (p^i, q^i) will typically be observations on production or consumption units that are separated spatially but yet are still comparable. For the sake of definiteness, we shall refer to the I data points as countries. Each commodity n is supposed to be the same across all countries. This can always be done by a suitable extension of the list of commodities.

Our first approach to the construction of a system of multilateral price and quantity indexes is based on the use of a bilateral quantity index Q. In this method, the first step is to pick the best bilateral index number formula, for example, the Fisher quantity index Q_F defined by (14) and (5) or the implicit Törnqvist-Theil quantity index QT defined by (22) and (5). Secondly, pick a numeraire country, say country 1, and then calculate the aggregate quantity for each country i relative to country 1 by evaluating the quantity index $Q(p^1, p^i, q^1, q^i)$. In order to put these relative quantity measures on a symmetric footing, we convert each relative to country 1 quantity measure into a share of world quantity by dividing through by $\sum_{k=1}^{I} Q(p^1, p^k, q^1, q^k)$. For a general numeraire country j, define the *share of world quantity for country i, using country j as the numeraire country*, by:

$$\sigma_i^j(p,q) \equiv Q(p^j, p^i, q^j, q^i) \Big/ \sum_{k=1}^{I} Q(p^j, p^k, q^j, q^k); \quad i = 1, \ldots, I, \tag{66}$$

where $p \equiv (p^1, \ldots, p^I)$ is the N by I matrix of price data and $q \equiv (q^1, \ldots, q^I)$ is the N by I matrix of quantity data. Once the numeraire country j has been chosen and the country i shares σ_i^j calculated, we may set $Q^i = \sigma_i^j$ and $p^i \equiv p^i \,.\, q^i/Q^i$ for $i = 1,\ldots, I$. Thus we have provided a solution to the multilateral index number problem (1). Of course, one is free to enormalize the resulting P^i and Q^i if desired: all Q^i can be multiplied by a number provided all P^i are divided by this same number. Kravis (1984) called this method the *star system*, since the numeraire country plays a starring role: all countries are compared with it and it alone.

Of course, the problem with the star system for making multilateral comparisons is its lack of invariance to the choice of the numeraire or star country. Different choices for the base country will in general give rise to different indexes P^i and Q^i. This problem can be traced to the lack of circularity of the bilateral formula Q: if Q satisfies the time reversal test and the circular test for quantity indexes, then $\sigma_i^j = \sigma_i^k$ for all i, j and k, that is, the shares σ_i^j defined by (66) do not depend on the choice of the numeraire country j. However, given that the chosen best bilateral formula does not satisfy the circularity test (as is the case with Q_F and Q_T), how can we generate multilateral indexes that treat each country symmetrically?

Fisher (1922, p. 305) recognized that the simplest way of achieving symmetry was to average base specific index numbers over all possible bases. Thus define country i's share of world output $S_i(p, q)$ by

$$S_j(p,q) \equiv \sum_{j=1}^{I} \sigma_i^j(p,q)/I, i = 1, \ldots, I \quad (67)$$

where the σ_i^j are defined by (66). We can now define country i quantities and prices by

$$Q^i \equiv S_i(p,q); P^i \equiv p^i \cdot q^i/Q^i, i = 1, \ldots, I \quad (68)$$

Fisher (1922, p. 305) called this method of constructing multilateral indexes the *blend method* while Diewert (1986) called it the *democratic weights method*, since each share of world output using each country as the base is given an equal weight in the formation of the average.

Of course, there is no need to use an arithmetic average of the σ_i^j as in (67); one can use a geometric average:

$$\sigma_i(p,q) \equiv \left[\prod_{j=1}^{I} \sigma_i^j(p,q)\right]^{1/I}, i = 1, \ldots, I. \quad (69)$$

Using (69), the resulting shares no longer sum to one in general, so country i's share of world output is now defined as:

$$S_i(p,q) \equiv \sigma_i(p,q)/\sum_{k=1}^{I} \sigma_k(p,q), i = 1, ..., I. \quad (70)$$

If the Fisher index Q_F is used in the definition of the σ_i^j, then

$$S_i(p,q)/S_j(p,q) = \left[\prod_{k=1}^{I} Q_F(p^k,p^i,q^k,q^i)/\prod_{m=1}^{I} Q_F(p^m,p^j,q^m,q^j)\right]^{1/I} \quad (71)$$

and in this case the multilateral method defined by (71) reduces to a method recommended by Gini (1924, 1931), Eltetö and Köves (1964) and Szulc (1964), the *GEKS method*. Instead of using the Fisher formula in (71), Caves et al. (1982a) advocated the use of the (direct) Törnqvist-Theil quantity index while Diewert (1986) suggested the use of the implicit translog quantity index Q_T defined by (5) when P is P_T defined by (22), since Q_T is well defined even in the case where some quantities q_n^i are negative. We call the indexes generated by (69) and (70) for a general bilateral index Q *generalized GEKS* indexes.

When forming averages of the σ_i^j as in (67) or (69), there is no necessity to use equal weights: one can define country j's value share of world output as $\beta_j \equiv p^j \cdot q^j/\sum_{k=1}^{I} p^k \cdot q^k$ (this requires all prices to be measured in units of a common currency) and then we may define a plutocratic share weighted average of the σ_i^j:

$$S_i(p,q) \equiv \sum_{j=1}^{I} \beta_j(p,q)\, \sigma_i^j(p,q). \quad (72)$$

Diewert (1986) called this method of constructing multilateral indexes the *plutocratic weights method*.

Another multilateral method that is based on a bilateral index Q may be described as follows. Define

$$\sigma_i(p,q) \equiv \sum_{j=1}^{I} \left[Q\left(p^j,p^i,q^j,q^i\right)^{-1}\right]^{-1}; \quad i = 1, \ldots, I. \quad (73)$$

If there is only one commodity so that $N = 1$ and the bilateral index Q satisfies quantity counterparts to tests T3 and T5, then $\alpha_i = \left[\sum_{j=1}^{I} (q_i/q^j)^{-1}\right]^{-1} = \left[\sum_{j=1}^{I} q^j/q^j\right]^{-1} = q^j/\sum_{j=1}^{I} q^j$ which is country i's share of world product. In the general case where $N > 1$, the 'shares' α_i do not necessarily sum up to unity, so it is necessary to normalize them:

$$S_i(p,q) \equiv \alpha_i(p,q)/\sum_{k=1}^{I} \alpha_k(p,q); \quad i = 1, \ldots, I. \quad (74)$$

Diewert (1986, 1988, 1999b) called this the *own share method* for making multilateral comparisons.

The above methods for achieving consistency and symmetry rely on averaging over various bilateral index number comparisons. Fisher (1922, p. 307) realized that symmetry could be achieved by making comparisons with an average, he called this broadening the base. Thus the *average basket method* (see Walsh 1901, p. 431; Gini 1931, p. 8; Fisher 1922, p. 307; Ruggles 1967; Diewert 1999b, pp. 24–5) may be described as follows. The price level of country I relative to country j is set equal to $p^i \cdot \left(\sum_{k=1}^{J} q^k/I\right)/p^j \cdot \left(\sum_{k=1}^{J} q^k/I\right)$. Now define $Q^{ji} \equiv [p^i \cdot q^i/p^j \cdot q^j]/[p^i \cdot (\sum_k q^k)/p^j \cdot (\sum_k q^k)]$ to be the implicit output of country i relative to j. Choose a j as a numeraire country and calculate country i's share of world output as:

$$S_i(p,q) \equiv Q^{ji} / \sum_{k=1}^{I} Q^{jk} = \left(p^i \cdot q^i / p^i \cdot \sum_k q^k\right) / \sum_{m=1}^{I} \left(p^m \cdot q^m / p^m \cdot \sum_k q^k\right); \quad i = 1, \ldots, I. \tag{75}$$

Note that the final expression for S_i does not depend on the choice of the numeraire country j. As usual, once the share functions, S_i, have been defined, the aggregate Q_i and P_i may be defined by (68).

A variation on the basket method due to Geary (1958) and Khamis (1972) is defined by (76–78) below:

$$\pi_n \equiv \sum_{i=1}^{I} p_n^i q_n^i / P^i \sum_{k=1}^{I} q_n^k, n = 1, \ldots, N; \tag{76}$$

$$p^i \equiv \sum_{n=1}^{N} p_n^i q_n^i / \sum_{m=1}^{N} \pi_m q_m^i, \; i = 1, \ldots, I; \tag{77}$$

$$Q_i \equiv p^i \cdot q^i / P^i, \; i = 1, \ldots, I. \tag{78}$$

π_n is interpreted as an average international price for good n. From (77), it can be seen that P^i, the price level or purchasing power parity for country i, is a Paasche- like price index for country i except that the base prices are chosen to be the international prices π_n. The π_n and $(P^i) - 1$ can be solved for as a system of simultaneous linear equations (up to a scalar normalization) or the $(P^i)^{-1}$ may be determined as the components of the eigenvector that corresponds to the maximal positive eigenvalue of a certain matrix. The P^i can be normalized so that the quantities Qi defined by (78) sum up to unity. This GK *method* for making multilateral comparisons has been widely used in empirical applications, for example, see Kravis et al. (1975).

We have defined seven methods for making multilateral comparisons: the star method (66), the democratic (67) and plutocratic (72) weights methods, the GEKS method (71), the own share method (74), the average basket method (75) and the GK method (78). Many additional methods have been suggested, for example, see Hill (1997), Diewert (1986, 1988, 1999b), Rao (1990), and Balk (1996). How can we discriminate among them? One helpful approach would be to define a system of *multilateral tests* and then evaluate how the above methods satisfy these tests. Space does not permit the development of this approach in this short survey, for applications of this approach, see Diewert (1988, 1999b) and Balk (1996). A clear consensus on the best multilateral method has not yet emerged.

We conclude this section by looking at a stochastic or descriptive statistics approach to making multilateral comparisons: namely, Summer's (1973) country product dummy (CPD) method for making multilateral comparisons. If there are I countries in the comparison and N products, the relationship of the prices between the various countries using the CPD model is given (approximately) by the following model:

$$p_n^c \approx \alpha_c \beta_n; \quad c = 1, \ldots, I; \quad n = 1, \ldots, N; \tag{79}$$

$$\alpha_1 = 1 \tag{80}$$

where p_n^c is the price (in domestic currency) of commodity n in country c. Quantities for each

commodity in each country are assumed to be measured in the same units. Equation 80 above is an identifying normalization, that is, we measure the price level of each country relative to the price level in country 1. Note that there are *IN* prices in the model and there are $I - 1 + N$ parameters to 'explain' these prices. Note also that the basic hypothesis that is implied by (79) is that commodity prices are approximately proportional between the two countries. Taking logarithms of both sides of (79) and adding error terms leads to the following CPD regression model:

$$\ln p_n^c = \ln \alpha_c + \ln \beta_n + \varepsilon_n^c; \quad c = 1, \ldots, I; \; n = 1, \ldots, N. \tag{81}$$

The main advantage of the CPD method for comparing prices across countries over traditional index number methods is that we can obtain *standard errors* for the country price levels $\alpha_2, \alpha_3, \ldots, \alpha_I$. This advantage of the stochastic approach to index number theory was stressed by Summers (1973) and more recently by Selvanathan and Rao (1994).

The recent literature on the CPD method notes that it is a special case of a hedonic regression model and this recent literature makes connections between weighted hedonic regressions and traditional index number formulae, see Triplett and McDonald (1977), Diewert (2003, 2005b, c, 2007b), de Haan (2004a, b), Silver (2003), and Silver and Heravi (2005).

Other Aspects of Index Number Theory

There are many important recent developments in index number theory that we cannot cover in any depth in this brief survey. Some of these developments are:

- Sampling problems and the construction of indexes at the first stage of aggregation: see Dalén (1992), Diewert (1995a), ILO (2004), and IMF (2004).
- *The treatment of seasonality*: see Turvey (1979), Balk (1980), (2005), Diewert (1983c), (1998b), (1999a), Hill (1996), Alterman et al. (1999), ILO (2004), and Armknecht and Diewert (2004).
- *The analysis of sources of bias in consumer price indexes*. This topic was greatly stimulated by the Boskin Commission Report, see Boskin et al. (1996). For additional contributions to this subject, see Diewert (1987, 1998a), Reinsdorf (1993), Schultze and Mackie (2002), Lebow and Rudd (2003), Balk and Diewert (2004), and ILO (2004).
- *Productivity indexes*. As more and more countries start programmes to measure sectoral and economy wide productivity, this topic has become more important. The original methodology for measuring productivity using index number techniques is due to Jorgenson and Griliches (1967, 1972) and it was first adopted by the U.S. Bureau of Labor Statistics (1983) and subsequently by Canada, Australia and more recently by New Zealand and Switzerland. Diewert (1976, 1983b), Caves et al. (1982b), Diewert and Morrison (1986), Kohli (1990), Morrison and Diewert (1990), Balk (1998, 2003), Schreyer (2001), Diewert and Fox (2004), Diewert and Nakamura (2003), and Diewert and Lawrence (2006) all made contributions connecting productivity measurement with index number theory.
- *Contribution analysis*. Suppose an aggregate price or quantity index shows a certain change over a certain period. Many analysts want to be able to compute the contribution of price or quantity change of specific components of the overall index and the problem of precisely defining such contributions has given rise to a fairly substantial recent literature. Contributors to this literature include Diewert (1983b, 2002a), Diewert and Morrison (1986), van IJzeren (1957, 1983, 1987), Kohli (1990, 2003, 2004, 2007), Morrison and Diewert (1990), Fox and Kohli (1998), and Reinsdorf et al. (2002).
- *Quality change*. The analysis thus far has assumed that the list of commodities in the aggregate is fixed and is unchanging and thus it is not able to deal with the problem of quality

change. For extensive discussions of this problem, see Triplett (2004) and the chapters on quality change in ILO (2004) and IMF (2004).

- *Index number theory in terms of differences rather than ratios*. Hicks (1941–42) noticed the similarities between measuring welfare change (difference measures) and index numbers of quantity change (ratio measures). The early literature on the difference approach dates back to Bennet (1920) and Montgomery (1929, 1937). More recent contributions to this subject may be found in Diewert (1992b, 2005a).

Since the mid-1980s interest in index number theory and economic measurement problems in general has increased. Perhaps influenced by Hill (1993), who in turn was influenced by Diewert (1976) and (1978), national statistical agencies are moving towards using chained superlative indexes as their target indexes: see Moulton and Seskin (1999) and Cage et al. (2003) for US developments. International agencies have also endorsed the use of superlative indexes as target indexes: see the manuals produced by the ILO (2004) and the IMF (2004). These manuals are a useful development since they help disseminate best practices and they help to harmonize statistics across countries, leading to a higher degree of accuracy and comparability. One hopes that these positive developments will continue.

Bibliography

Afriat, S.N. 1972. The theory of international comparisons of real income and prices. In *International comparisons of prices and outputs*, ed. D.J. Daley. New York: Columbia University Press.

Allen, R.G.D. 1949. The economic theory of index numbers. *Economica* NS 16: 197–203.

Alterman, W.F., W.E. Diewert, and R.C. Feenstra. 1999. *International trade price indexes and seasonal commodities*. Washington, DC: Bureau of Labor Statistics.

Armknecht, P.A., and W.E. Diewert. 2004. Treatment of seasonal products. In *Producer price index manual: Theory and practice*. Washington, DC: International Monetary Fund.

Balk, B.M. 1980. A method for constructing price indices for seasonal commodities. *Journal of the Royal Statistical Society, Series A* 143: 68–75.

Balk, B.M. 1985. A simple characterization of Fisher's price index. *Statistische Hefte* 26: 59–63.

Balk, B.M. 1995. Axiomatic price index theory: A survey. *International Statistical Review* 63: 69–93.

Balk, B.M. 1996. A comparison of ten methods for multilateral international price and volume comparisons. *Journal of Official Statistics* 12: 199–222.

Balk, B.M. 1998. *Industrial price, quantity and productivity indices*. Boston: Kluwer.

Balk, B.M. 2003. The residual: On monitoring and benchmarking firms, industries, and economies with respect to productivity. *Journal of Productivity Analysis* 20: 5–47.

Balk, B.M. 2005. Annual and quarterly productivity measures. Paper presented at the Economic Measurement Group Workshop, Coogee, 12–13 Dec.

Balk, B.M., and W.E. Diewert. 2001. A characterization of the Törnqvist priceindex. *Economics Letters* 73: 279–281.

Balk, B.M., and W.E. Diewert. 2004. The Lowe consumer price index and its substitution bias. Discussion Paper No. 04–07. Department of Economics, University of British Columbia.

Bennet, T.L. 1920. The theory of measurement of changes in cost of living. *Journal of Royal Statistical Society* 83: 455–462.

Boskin, M.J. (chair), E.R. Dullberger, R.J. Gordon, Z. Griliches, and D.W. Jorgenson. 1996. *Final report of the commission to study the consumer price index*. US Senate, Committee on Finance. Washington, DC: US Government Printing Office.

Bowley, A.L. 1899. Wages, nominal and real. In *Dictionary of Political Economy*, ed. R.H.L. Palgrave, vol. vol. 3. London: Macmillan.

Bowley, A.L. 1901. *Elements of statistics*. Westminster: P.S. King and Son.

Bowley, A.L. 1919. The measurement of changes in the cost of living. *Journal of the Royal Statistical Society* 82: 343–372.

Bowley, A.L. 1928. Notes on index numbers. *Economic Journal* 38: 216–237.

Cage, R., J.S. Greenlees, and P. Jackman. 2003. Introducing the chained CPI. Paper presented at the Seventh Meeting of the International Working Group on Price Indices, (Ottawa Group), Paris.

Carli, G.-R. 1764. Del valore e dellaproporzione de' metallimonetati. In *Scrittoriclassiciitaliani di economiapolitica*, vol. 13. Milano: G.G. Destefanis, 1804.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982a. Multilateral comparisons of output, input and productivity using superlative index numbers. *Economic Journal* 92: 73–86.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982b. The economic theory of index numbers and the measurement of input, output and productivity. *Econometrica* 50: 1393–1414.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1971. Conjugate duality and the transcendental logarithmic production function. *Econometrica* 39: 255–256.

Clements, K.W., H.Y. Izan, and E.A. Selvanathan. 2006. Stochastic index numbers: A review. *International Statistical Review* 74: 235–270.

Dalén, J. 1992. Computing elementary aggregates in the Swedish consumer price index. *Journal of Official Statistics* 8: 129–147.

de Haan, J. 2004a. Direct and indirect time dummy approaches to hedonic price measurement. *Journal of Economic and Social Measurement* 29: 427–443.

de Haan, J. 2004b. Hedonic regression: The time dummy index as a special case of the imputation Törnqvist index. Paper presented at the 8th Ottawa Group Meeting, Helsinki.

Diewert, W.E. 1974. Applications of duality theory. In *Frontiers of quantitative economics*, ed. M.-D. Intriligator and D.A. Kendrick, vol. 2. Amsterdam: North-Holland.

Diewert, W.E. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4: 114–145.

Diewert, W.E. 1978. Superlative index numbers and consistency in aggregation. *Econometrica* 46: 883–900.

Diewert, W.E. 1981. The economic theory of index numbers: A survey. In *Essays in the theory and measurement of consumer behaviour in honour of Sir Richard Stone*, ed. A. Deaton. London: Cambridge University Press.

Diewert, W.E. 1983a. The theory of the cost of living index and the measurement of welfare change. In *Price level measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Diewert, W.E. 1983b. The theory of the output price index and the measurement of real output change. In *Price level measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Diewert, W.E. 1983c. The treatment of seasonality in a cost of living index. In *Price level measurement*, ed. W.-E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Diewert, W.E. 1986. *Microeconomic approaches to the theory of international comparisons*. Technical Working Paper No. 53. Cambridge, MA: NBER.

Diewert, W.E. 1987. Index numbers. In *The New Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London: Macmillan.

Diewert, W.E. 1988. Test approaches to international comparisons. In *Measurement in economics*, ed. W. Eichhorn. Heidelberg: Physica-Verlag.

Diewert, W.E. 1992a. Fisher ideal output, input and productivity indexes revisited. *Journal of Productivity Analysis* 3: 211–248.

Diewert, W.E. 1992b. Exact and superlative welfare change indicators. *Economic Inquiry* 30: 565–582.

Diewert, W.E. 1993a. The early history of price index research. In *Essays in index number theory*, ed. W.-E. Diewert and A.O. Nakamura. Amsterdam: North-Holland.

Diewert, W.E. 1993b. Symmetric means and choice under uncertainty. In *Essays in index number theory*, ed. W.-E. Diewert and A.O. Nakamura. Amsterdam: North-Holland.

Diewert, W.E. 1995a. Axiomatic and economic approaches to elementary price indexes. Discussion Paper No. 95-01. Department of Economics, University of British Columbia.

Diewert, W.E. 1995b. On the stochastic approach to index numbers. Discussion Paper No. 95-31. Department of Economics, University of British Columbia.

Diewert, W.E. 1997. Commentary on Mathew D. Shapiro and David W. Wilcox: Alternative strategies for aggregating price in the CPI. *Federal Reserve Bank of St. Louis Review* 79: 127–137.

Diewert, W.E. 1998a. Index number issues in the consumer price index. *Journal of Economic Perspectives* 12 (1): 47–58.

Diewert, W.E. 1998b. High inflation, seasonal commodities and annual index numbers. *Macroeconomic Dynamics* 2: 456–471.

Diewert, W.E. 1999a. Index number approaches to seasonal adjustment. *Macroeconomic Dynamics* 3: 48–68.

Diewert, W.E. 1999b. Axiomatic and economic approaches to multilateral comparisons. In *International and interarea comparisons of income, output and prices*, ed. A. Heston and R.E. Lipsey. Chicago: University of Chicago Press.

Diewert, W.E. 2001. The consumer price index and index number purpose. *Journal of Economic and Social Measurement* 27: 167–248.

Diewert, W.E. 2002a. The quadratic approximation lemma and decompositions of superlative indexes. *Journal of Economic and Social Measurement* 28: 63–88.

Diewert, W.E. 2002b. Harmonized indexes of consumer prices: Their conceptual foundations. *Swiss Journal of Economics and Statistics* 138: 547–637.

Diewert, W.E. 2003. Hedonic regressions: a review of some unresolved issues. Paper presented at the 7th Meeting of the Ottawa Group, Paris.

Diewert, W.E. 2004. A new axiomatic approach to index number theory. Discussion Paper No. 04-05. Department of Economics, University of British Columbia.

Diewert, W.E. 2005a. Index number theory using differences instead of ratios. *American Journal of Economics and Sociology* 64: 311–360.

Diewert, W.E. 2005b. Weighted country product dummy variable regressions and index number formulae. *Review of Income and Wealth* 51: 561–571.

Diewert, W.E. 2005c. Adjacent period dummy variable hedonic regressions and bilateral index number theory. Discussion Paper No. 05-11. Department of Economics, University of British Columbia.

Diewert, W.E. 2007a. Similarity indexes and criteria for spatial linking. In *Purchasing power parities of currencies: Recent advances in methods and applications*, ed. D.S. Prasada Rao. Cheltenham: Edward Elgar.

Diewert, W.E. 2007b. On the stochastic approach to linking the regions in the ICP. In *Price and productivity measurement*, ed. W.E. Diewert et al. Vancouver: Trafford Press.

Diewert, W.E., and K.J. Fox. 2004. On the estimation of returns to scale, technical progress and monopolistic

markups. Discussion Paper No. 04-09. Department of Economics, University of British Columbia.
Diewert, W.E., and D. Lawrence. 2006. *Measuring the contributions of productivity and terms of trade to Australia's economic welfare*. Report by Meyrick and Associates to the Australian Government. Canberra: Productivity Commission.
Diewert, W.E., and C.J. Morrison. 1986. Adjusting output and productivity indexes for changes in the terms of trade. *Economic Journal* 96: 659–679.
Diewert, W.E., and A.O. Nakamura. 2003. Index number concepts, measures and decompositions of productivity growth. *Journal of Productivity Analysis* 19: 127–159.
Edgeworth, F.Y. 1888. Some new methods of measuring variation in general prices. *Journal of the Royal Statistical Society* 51: 346–368.
Edgeworth, F.Y. 1923. The doctrine of index numbers according to Mr. Correa Walsh. *Economic Journal* 33: 343–351.
Edgeworth, F.Y. 1925. *Papers relating to political economy*. Vol. 1. New York: Burt Franklin.
Eichhorn, W. 1978. *Functional equations in economics*. London: Addison-Wesley.
Eichhorn, W., and J. Voeller. 1976. *Theory of the price index*. Berlin: Springer.
Eltetö, O., and P. Köves. 1964. On a problem of index number computation relating to international comparison. *Statisztikai Szemle* 42: 507–518.
Fisher, I. 1911. *The purchasing power of money*. London: Macmillan.
Fisher, W.C. 1913. The tabular standard in Massachusetts history. *Quarterly Journal of Economics* 27: 417–451.
Fisher, I. 1921. The best form of index number. *Journal of the American Statistical Association* 17: 535–537.
Fisher, I. 1922. *The making of index numbers*. Boston: Houghton Mifflin.
Fox, K.J., and U. Kohli. 1998. GDP growth, terms of trade effects and total factor productivity. *The Journal of International Trade and Economic Development* 7: 87–110.
Frisch, R. 1936. Annual survey of economic theory: The problem of index numbers. *Econometrica* 4: 1–39.
Funke, H., and J. Voeller. 1978. A note on the characterization of Fisher's ideal index. In *Theory and applications of economic indices*, ed. W. Eichhorn et al. Würzburg: Physica-Verlag.
Funke, H., G. Hacker, and J. Voeller. 1979. Fisher's circular test reconsidered. *Schweizerische Zeitschrift für Volkswirtshaft und Statistik* 115: 677–687.
Geary, R.G. 1958. A note on comparisons of exchange rates and purchasing power between countries. *Journal of the Royal Statistical Society Series A* 121: 97–99.
Gini, C. 1924. Quelques considérations au sujet de la construction des nombres indices des prix et des questions analogues. *Metron* 4: 3–162.
Gini, C. 1931. On the circular test of index numbers. *Metron* 9: 3–24.
Hicks, J.R. 1941–42. Consumers' surplus and index numbers. *Review of Economic Studies* 9, 126–137.
Hill, T.P. 1988. Recent developments in index number theory and practice. *OECD Economic Studies* 10: 123–148.
Hill, T.P. 1993. Price and volume measures. In *System of national accounts 1993*. Luxembourg/Paris/New York/-Washington, DC: Eurostat/IMF/OECD/UN/World Bank.
Hill, T.P. 1996. *Inflation accounting*. Paris: OECD.
Hill, R.J. 1997. A taxonomy of multilateral methods for making international comparisons of prices and quantities. *Review of Income and Wealth* 43: 49–69.
Hill, R.J. 1999a. Comparing price levels across countries using minimum spanning trees. *The Review of Economics and Statistics* 81: 135–142.
Hill, R.J. 1999b. International comparisons using spanning trees. In *International and interarea comparisons of income, output and prices*, Studies in income and wealth, NBER, ed. A. Heston and R.E. Lipsey, vol. 61. Chicago: University of Chicago Press.
Hill, R.J. 2001. Measuring inflation and growth using spanning trees. *International Economic Review* 42: 167–185.
Hill, R.J. 2004. Constructing price indexes across space and time: The case of the European Union. *American Economic Review* 94: 1379–1410.
Hill, R.J. 2006a. Superlative index numbers: Not all of them are super. *Journal of Econometrics* 130: 25–43.
Hill, R.J. 2006b. When does chaining reduce the Paasche-Laspeyres spread? An application to scanner data. *Review of Income and Wealth* 52: 309–329.
Hill, R.J. 2007. Comparing per capita income levels across countries using spanning trees: Robustness, prior restrictions, hybrids and hierarchies. In *Purchasing power parities of currencies: Recent advances in methods and applications*, ed. D.S. Prasada Rao. Cheltenham: Edward Elgar.
ILO/IMF/OECD/UNECE/Eurostat/World Bank. 2004. In *Consumer price index manual: Theory and practice*, ed. P. Hill. Geneva: International Labour Office.
IMF/ILO/OECD/UNECE/Eurostat/World Bank. 2004. In *Producer price index manual: Theory and practice*, ed. P. Armknecht. Washington, DC: International Monetary Fund.
Jevons, W.S. 1865. The variation of prices and the value of the currency since 1782. *Journal of the Statistical Society of London* 28: 294–320. Reprinted in *Investigations in currency and finance*. London: Macmillan and Co, 1884.
Jevons, W.S. 1884. A serious fall in the value of gold ascertained and its social effects set forth (1863). In *Investigations in Currency and Finance*. London: Macmillan and Co.
Jorgenson, D.W., and Z. Griliches. 1967. The explanation of productivity change. *Review of Economic Studies* 34: 249–283.
Jorgenson, D.W., and Z. Griliches. 1972. Issues in growth accounting: A reply to Edward F. Denison. *Survey of Current Business* 52: 65–94.
Keynes, J.M. 1930. *Treatise on money*. Vol. 1. London: Macmillan.

Khamis, S.H. 1972. A new system of index numbers for national and international purposes. *Journal of the Royal Statistical Society, Series A* 135: 96–121.

Kohli, U. 1990. Growth accounting in the open economy: Parametric and nonparametric estimates. *Journal of Economic and Social Measurement* 16: 125–136.

Kohli, U. 2003. Growth accounting in the open economy: International comparisons. *International Review of Economics and Finance* 12: 417–435.

Kohli, U. 2004. Real GDP, real domestic income and terms of tradechanges. *Journal of International Economics* 62: 83–106.

Kohli, U. 2007. Terms of trade, real exchange rates, and trading gains. In *Price and productivity measurement*, ed. W.E. Diewert et al. Vancouver: Trafford Press.

Konüs, A.A. 1924. The problem of the true index of the cost of living. Trans. in *Econometrica* 7 (1939), 10–29.

Konüs, A.A., and S.S. Byushgens. 1926. K probleme pokupatelnoi cili deneg. *Voprosi konyunkturi* 2: 151–172.

Kravis, I.B. 1984. Comparative studies of national incomes and prices. *Journal of Economic Literature* 22: 1–39.

Kravis, I.B., Z. Kenessey, A. Heston, and R. Summers. 1975. *A system of international comparisons of gross product and purchasing power*. Baltimore: Johns Hopkins University Press.

Laspeyres, E. 1871. Die berechnung einer mittleren waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik* 16: 296–314.

Lebow, D.E., and J.B. Rudd. 2003. Measurement error in the consumer price index: Where do we stand? *Journal of Economic Literature* 41: 159–201.

Lehr, J. 1885. *Beitrage zur Statistik der Preise*. Frankfurt: J.D. Sauerlander.

Lowe, J. 1823. *The present State of England in regard to agriculture, trade and finance*. 2nd ed. London: Longman, Hurst, Rees, Orme and Brown.

Malmquist, S. 1953. Index numbers and indifference surfaces. *Trabajos de Estadística* 4: 209–242.

Marshall, A. 1887. Remedies for fluctuations of general prices. *Contemporary Review* 51: 355–375.

Montgomery, J.K. 1929. *Is there a theoretically correct price index of a group of commodities?* Rome: Roma L'Universale Tipogr. Poliglotta (privately printed).

Montgomery, J.K. 1937. *The mathematical problem of the price index*. Orchard House: P.S. King & Son.

Morrison, C.J., and W.E. Diewert. 1990. Productivity growth and changes in the terms of trade in Japan and the United States. In *Productivity growth in Japan and the United States*, ed. C.R. Hulten. Chicago: University of Chicago Press.

Moulton, B.R., and E.P. Seskin. 1999. A preview of the 1999 comprehensive revision of the national income and product accounts. *Survey of Current Business* 79: 6–17.

Paasche, H. 1874. Über die preisentwicklung der letzten Jahre nach den hamburger borsennotirungen. *Jahrbücher für Nationalökonomie und Statistik* 12: 168–178.

Pierson, N.G. 1895. Index numbers and appreciation of gold. *Economic Journal* 5: 329–335.

Pierson, N.G. 1896. Further considerations on index-numbers. *Economic Journal* 6: 127–131.

Pollak, R.A. 1980. Group cost-of-living indexes. *American Economic Review* 70: 273–278.

Pollak, R.A. 1981. The social cost-of-living index. *Journal of Public Economics* 15: 311–336.

Pollak, R.A. 1983. The theory of the cost-of-living index. In *Price level measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Rao, D.S.P. 1990. A system of log-change index numbers for multilateral comparisons. In *Comparisons of prices and real products in Latin America*, ed. J. Salazar-Carillo and D.S. PrasadaRao. New York: Elsevier.

Reinsdorf, M. 1993. The effect of outlet price differentials on the U.S. consumer price index. In *Price measurement and their uses*, ed. M.F. Foss, M.E. Manser, and A.H. Young. Chicago: University of Chicago Press.

Reinsdorf, M. 2007. Axiomatic price index theory. In *Measurement in economics: A handbook*, ed. M.J. Boumans. Amsterdam: Elsevier.

Reinsdorf, M., and A. Dorfman. 1999. The monotonicity axiom and the Sato-Vartia index. *Journal of Econometrics* 90: 45–61.

Reinsdorf, M.B., W.E. Diewert, and C. Ehemann. 2002. Additive decompositions for the fisher, Törnqvist and geometric mean indexes. *Journal of Economic and Social Measurement* 28: 51–61.

Ruggles, R. 1967. Price indexes and international price comparisons. In *Ten economic studies in the tradition of Irving Fisher*, ed. W. Fellner. New York: Wiley.

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

Samuelson, P.A., and S. Swamy. 1974. Invariant economic index numbers and canonical duality: Survey and synthesis. *American Economic Review* 64: 566–593.

Schreyer, P. 2001. *OECD productivity manual: A guide to the measurement of industry-level and aggregate productivity growth*. Paris: OECD.

Schultze, C.L., and C. Mackie. 2002. *At what price? Conceptualizing and measuring cost-of living and price indices*. Washington, DC: National Academy Press.

Scrope, G.P. 1833. *Principles of political economy*. London: Longman, Rees, Orme, Brown, Green and Longman.

Selvanathan, E.A., and D.S. PrasadaRao. 1994. *Index numbers: A stochastic approach*. Ann Arbor: University of Michigan Press.

Shephard, R.W. 1953. *Cost and production functions*. Princeton: Princeton University Press.

Sidgwick, H. 1883. *The principles of political economy*. London: Macmillan.

Silver, M. 2003. The use of weights in hedonic regressions: The measurement of quality adjusted price changes. Room document for the 7th meeting of the Ottawa Group, Paris.

Silver, M., and S. Heravi. 2005. A failure in the measurement of inflation: Results from a hedonic and matched experiment using scanner data. *Journal of Business and Economic Statistics* 23: 269–281.
Summers, R. 1973. International comparisons with incomplete data. *Review of Income and Wealth* 29: 1–16.
Szulc, B. 1964. Indices for multiregional comparisons. *Przeglad Statystyczny* 3: 239–254.
Szulc, B.J. 1983. Linking price index numbers. In *Price level measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.
Theil, H. 1967. *Economics and information theory.* Amsterdam: North-Holland.
Törnqvist, L. 1936. The bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin* 10: 1–8.
Törnqvist, L., and E. Törnqvist. 1937.Vilket är förhällandet mellan finska markens och svenska kronans köpkraft? *Ekonomiska Samfundets Tidskrift* 39, 1–39. Repr. In *Collected Scientific Papers of Leo Törnqvist.* Helsinki: Research Institute of the Finnish Economy, 1981.
Triplett, J.E. 2004. *Handbook on hedonic indexes and quality adjustments in price indexes: Special application to information technology products*. Working Paper 2004/9. Paris: Directorate for Science, Technology and Industry, OECD.
Triplett, J.E., and R.J. McDonald. 1977. Assessing the quality error in output measures: the case of refrigerators. *Review of Income and Wealth* 23: 137–156.
Turvey, R. 1979. *The treatment of seasonal items in consumer price indices*, Bulletin of labour statistics, fourth quarter, 13–33. Geneva: International Labour Office.
US Bureau of Labor Statistics. 1983. *Trends in multifactor productivity, 1948–81*, Bulletin 2178. Washington, DC: U.S. Government Printing Office.
van IJzeren, J. 1957. *Three methods of comparing the purchasing power of currencies*, Statistical studies. Vol. 7. The Hague: Central Bureau of Statistics.
van IJzeren, J. 1983. *Index numbers for binary and multilateral comparisons. Statistical studies no. 34*. The Hague: Central Bureau of Statistics.
van IJzeren, J. 1987. *Bias in international index numbers: A mathematical elucidation*. Dissertation for the Hungarian Academy of Sciences. The Hague: Koninklijke Bibliotheek.
Vartia, Y.O. 1976. *Relative changes and index numbers*. Helsinki: Research Institute of the Finnish Economy.
Vogt, A. 1980. Der Zeit und der Faktorumkehrtestals 'Finders of Tests'. *StatistischeHefte* 21: 66–71.
Vogt, A., and J. Barta. 1997. *The making of tests for index numbers*. Heidelberg: Physica-Verlag.
von der Lippe, P. 2001. *Chain indices: A study in price index theory, spectrum of federal statistics*. Vol. 16. Wiesbaden: Statistisches Bundesamt.
Walsh, C.M. 1901. *The measurement of general exchange value*. New York: Macmillan and Co..
Walsh, C.M. 1921a. *The problem of estimation*. London: P.S. King &Son.
Walsh, C.M. 1921b. Discussion. *Journal of the American Statistical Association* 17: 537–544.
Westergaard, H. 1890. *Die Grundzüge der Theorie der Statistik*. Jena: Fischer.
Wynne, M.A. 1997. Commentary on measuring short run inflation for central bankers. *Federal Reserve Bank of St. Louis Review* 79: 161–167.
Young, A. 1812. *An inquiry into the progressive value of money in England as marked by the price of agricultural products*. London: Macmillan.

Indexed Securities

Alicia H. Munnell and Joseph B. Grolnic

Conventional securities are generally offered at a fixed coupon rate that incorporates the underlying expected real rate of return in the economy, the market's expectation at the time the security is issued of inflation over the duration of the instrument, a premium to compensate for the fact that future rates of inflation are uncertain, and an adjustment reflecting the tax treatment of interest on behalf of both the lender and the borrower. For simplicity, it is useful to abstract temporarily from the inflation risk premium and taxes, although both these factors will be discussed later.

With these simplifying assumptions, if the real rate of return in the economy is 3 per cent, and inflation is expected to remain constant at 4 per cent annually, the nominal return will be 7 per cent. If expectations should prove incorrect and inflation turns out to be lower than anticipated, say 2 per cent, investors will receive more income in present value terms than they expected and experience an increase in their real rate of return, reflecting the unanticipated decline in the inflation rate. On the other hand, if inflation turns out to be 6 per cent, then investors will receive less in real terms than expected and their real return will fall below the rate initially negotiated. If they attempt to sell the security in the higher inflationary environment, they will experience a capital loss.

Index bonds are financial instruments designed to protect investors fully against the

erosion of principal and interest due to inflation. This protection is accomplished in one of two ways. Under the first option, the bond is issued at a specified real coupon rate and both coupon payment and repayment of principal are scaled up or down by the change in prices that occurs between the time that the money is borrowed and the time the payments are made. For example, if inflation is 4 per cent annually, the coupon on a five-year $1000 bond issued at a real interest rate of 3 per cent would increase from $30 in the first year to $35.10 in the fifth year. At maturity, the government then would adjust the principal for inflation over the life of the bond; thus, in the above example, the government would repay $1217 at the end of the five-year period. This approach is similar to the index bonds that have been sold in Great Britain.

Under the second approach the entire inflation adjustment is made through the coupon payment and the bondholder is repaid his original principal at maturity. For example, if the real rate is set at 3 per cent and inflation averages 4 per cent, the total annual interest cost would be 7 per cent. This approach mimics the current method of compensating the lender for inflation, except that instead of trying to predict inflation at the time of the loan and incorporating this expectation into the stated nominal interest rate, actual observations on price are used to determine annual interest payments.

Either of these two approaches will protect the investor against the risks associated with unanticipated price changes if the index bond is held to maturity; however, it is important to emphasize that neither produces a risk-free investment. As with any long-term security, bondholders selling an index bond before maturity would take a capital loss if the underlying expected real rates have increased since the date of purchase. The result is that these bonds would probably not be the ideal assets for individuals to purchase directly unless they were certain that they could hold them to maturity. For index bonds to serve as risk-free inflationprotected investments, financial intermediaries are required which will hold the bonds to maturity and offer repackaged investments free of the real-return risk.

Impact on the Government Budget

Arguments about the potential impact of index bonds on the government budget have figured prominently in debates about this form of financing and the range of opinion has been extraordinary. In Great Britain, some opponents argued that index bonds would cost the government more than fixed-interest securities to service, since they would have to be issued at positive real interest rates as opposed to the negative real returns received by investors on nominal debt during the period 1973–8 (Rutherford 1983). On the other hand, in hearings before the Joint Economic Committee in May 1985, a major proponent of index bonds projected that, because excessive inflation premiums were incorporated in current yields, the US government could save $9 billion in the first year and $135 billion over a five-year period by issuing indexed rather than conventional long-term debt (Joint Economic Committee 1985).

These conflicting statements are based on opposite assumptions about people's ability to project future inflation. The contention that index bonds will cost money assumes that individuals will continually underestimate future inflation and always end up with lower than anticipated or negative returns; the argument that the Treasury can reduce costs with indexed debt assumes individuals will consistently overestimate inflation and demand excessive inflation premiums. It is unclear why, over the long run, individuals should systematically err on one side or another in their inflation projections.

In the last 15 years, the relationship between expected and actual inflation has varied over time; during the 1970s average expectations about near-term inflation tended to prove too low, while since 1981 inflation has generally fallen one or two percentage points below projections. Although no evidence is available on investors' ability to forecast inflation over longer periods, of say 20 or 30 years, the same pattern is likely to emerge as swings in short-run expectations affect the longer-run outlook. Hence, the most reasonable conclusion is that in the long run forecasting errors will cancel out, and have little impact on the relative costs of indexed versus unindexed debt.

On the other hand, the uncertainty surrounding future rates of inflation means that investors demand an inflation-risk premium before they are willing to take on fixed-coupon debt. In this case, the guarantee of a real return provided by indexed securities, which eliminates the risk of reduced real returns and capital losses caused by unanticipated inflation, would lower the yield that lenders will require in order to provide their funds. In other words, the lender would be willing to accept a somewhat lower rate in return for the privilege of having the government guarantee the real return on the loan.

Little evidence exists about the size of the inflation risk premium (an exception is Bodie et al. 1986). As long as the outlook for price increases is moderate, the premium is probably relatively small; at higher and more volatile rates of inflation, the importance of risk protection would increase. Even if this premium proved to be quite small, however, its elimination could produce substantial savings in view of the enormous magnitude of government debt. The problem is that, in the short run at least, the risk premium effect is likely to be dominated by the difference between expected and actual real returns caused by errors in investors' expectations. Hence, for any defined period of time, it would be impossible to predict whether substituting index bonds for traditional government securities would cost or save the Treasury money. In the long run, however, if errors in inflation forecasts cancel out, index bonds should save the government the inflation-risk premium on long-term securities.

While the net interest saving to the government is difficult to predict, the pattern of government borrowing would certainly be altered if the British indexing option were adopted. Even in an environment where inflationary expectations always prove correct and the inflation premium is zero, an index bond that defers the principal adjustment for inflation until maturity reduces the Treasury's borrowing in the intervening years.

Tax Policy and Index Bonds

Uncertainty about how index bonds would be taxed has been viewed as a major impediment to their introduction. The tax questions are indeed critical, because they determine not only how index bonds would affect revenues but also who might be the likely buyers of these securities and the potential yields.

If the tax code does not distinguish between real and inflationary returns, then the most likely results would be to tax both the real component of interest and the inflation adjustment as ordinary income. This would be quite straightforward in the case of the indexing method that incorporates the inflation adjustment in the interest rate, but some complexities arise in the case of the British approach. In order to make the treatment of bonds indexed in this fashion analogous to that accorded conventional and zero-coupon bonds, the annual appreciation of principal due to inflation would have to be taxed as it accrued.

Taxing the principal adjustment as if it were received each year would make index bonds less attractive than their unindexed counterparts. Not only would owners of securities have to pay taxes on illusory gains, which they do in the case of conventional bonds, but they would also have to pay the tax before they received their inflation compensation. On the other hand, deferring the tax on the adjustment of principal until the bond is redeemed at maturity would favour the indexed over the unindexed security and result in a loss of revenue for the Treasury.

The second problem with applying current tax law to index bonds is that it would no longer be possible to guarantee a constant real after-tax rate of return. Under the current system, taxes would rise with inflation and the real after-tax return would decline. For example, if the tax rate were 30 per cent, the real return on a bond with a 3 per cent coupon would be 2.1 per cent in an environment of no inflation. If inflation should rise to 4 per cent and the nominal coupon rises to only 7 per cent, the after-tax yield is 4.9 per cent or 0.9 per cent real. The only way to avoid this problem is to exempt the nominal adjustments for inflation from taxation. This approach, however, would introduce a type of inflation indexing not found elsewhere in the tax system.

Private Issues of Index Bonds

Some sceptics charge that if index bonds were such a great idea, they would have been offered by the private sector. Indeed, theoretical work by Stanley Fischer leads to the conclusion that firms should be equally willing to issue index bonds as conventional nominal bonds (Fischer 1982). Fischer offers two possible reasons for the lack of the private sector innovation: the relatively stable rates of inflation traditionally experienced in the United States and the possibility that borrowers' expectations about inflation have been systematically higher than those of lenders. Others contend that the issuance of index-linked debt may actually have been illegal in the United States until 1977 (McCulloch 1980). Another problem is that an aggregate price index may not correlate with prices received by an individual firm. The most persuasive reason, however, relates to the lack of indexation in the corporate income tax, which causes the effective tax rate to increase with inflation. If firms were to issue index bonds, this inverse relation between inflation and profitability would worsen, since corporations would forfeit the mitigating effect of the decline in the value of outstanding liabilities as inflation increased. Hence, the non-issuance of index bonds by the corporate sector may be one of the major casualties of an unindexed tax structure.

The only serious objection ever levelled against index bonds is that protecting bondholders from inflation might reduce public pressure to maintain price stability. If part of the pain of inflation is removed, this reasoning goes, the public's resolve to control inflation will weaken, and inflation will ultimately get worse. On the other hand, one could argue in economic terms that index bonds might help in the fight against inflation by providing an attractive investment vehicle that would encourage saving and, as argued by Tobin, by offering the monetary authorities a tool that would strengthen their control of the economy (Tobin 1971). In political terms, it would seem that the issuance of index bonds would eliminate one of the main incentives for the government to inflate the economy. With indexed debt the government can no longer reduce the real value of its outstanding liabilities by allowing prices to rise; instead, inflation will produce an immediate increase in required expenditures. Finally, index bonds do not appear to have encouraged inflation in Great Britain; the inflation rate has declined from 15 to 5 per cent since 1981, the year the bonds were introduced.

See Also

- ▶ Inflation
- ▶ Monetary Policy
- ▶ Wage Indexation

Bibliography

This essay is abstracted from Munnell and Grolnic (1986).

Bodie, Z., A. Kane, and R. McDonald. 1986. Risk and required returns on debt and equity. In *Financing corporate capital formation*, ed. B.M. Friedman, 51–66. New York: National Bureau of Economic Research.

Fischer, S. 1982. On the nonexistence of privately issued index bonds in the US capital market. In *Inflation, debt, and indexation*, ed. R. Dornbusch and M.H. Simonsen, 247–266. Cambridge, MA: MIT Press.

Joint Economic Committee. 1985. *Inflation indexing of government securities*. Hearing before the subcommittee on trade, productivity, and economic growth, 99 congress, 1 session, 14 May.

McCulloch, J.H. 1980. The ban on indexed bonds, 1933–77. *American Economic Review* 70: 1018–1021.

Munnell, A.H., and J.B. Grolnic. 1986. Should the US Government issue index bonds? Federal Reserve Bank of Boston. *New England Economic Review* 3–21.

Rutherford, J. 1983. Index-linked gilts. *National Westminster Review* 2–17.

Tobin, J. 1971. An essay on the principles of debt management. In *Essays in economics, volume 1: Macroeconomics*, ed. J. Tobin, 439–447. Chicago: Markham.

India, Economics in

Deepak Lal

Abstract

This article outlines the debates amongst Indian economists on planning, transforming agriculture, poverty and income distribution, and political economy and institutions. It

shows that much of this work pioneered many analyses which have come to define the sub-discipline of 'development economics'.

Keywords

Agricultural subsidies; Agriculture and economic development; Attached labour; Bauer, P; Bhagwati, J; Caste system; Central planning; Comparative advantage; Conjunctural poverty; Delhi School; Destitution; Development economics; Dirigisme; Dual economies; Dutch disease; Exchange controls; Famine; Farm size; Fel'dman, G. A; Food security; Free trade; Friedman, M; Green Revolution; Homo economicus; Import substitution; India, economics in; Industrialization; Johnson, H. G; Land reform; Leisure; Lewis, W. A; Mahalanobis, P. C; Malthus's theory of population; Meade, J. E; Pastoralism; Peasants; Population growth; Poverty; Poverty alleviation; Price controls; Privatization; Protection; Public works; Ramaswami, V. K; Redistribution of income; Rent seeking; Rural employment; Sen, A; Shadow prices; Sharecropping; Shenoy, B. R; Stalin, J; Structural poverty; Surplus labour; Terms of trade; Usury

JEL Classification

B2

Economics in India has been mainly concerned with finding means to alleviate its ancient and pervasive poverty. In this article I will concentrate on the debates amongst Indian economists, highlighting the contributions they have made in the process to the new discipline of 'development economics'.

The Indian economic debate began in the early twentieth century when after nearly a century of British colonial rule there were few signs of poverty alleviation, with only a modest rise in per capita income over the period (Sivasubramonian 2000). A nationalist and Marxist literature evolved, which laid the blame for this economic stagnation on alien rule and the implementation – since the 1850s – of the twin classical liberal principles (dominant in the metropolitan centre) of laissez-faire and 'free trade'. Alien rule was epitomized by the fiscal drain of resources from India to Britain (Naoroji 1901; Dutt 1904). Free trade was held responsible for India's failure to industrialize and the destruction of its extensive pre-colonial handloom textile industry.

By the 1930s, the Great Depression and Stalinist Russia's success in rapidly industrializing a large, poor and mainly agrarian economy coloured the thinking of Indian economists and political leaders like Nehru. A series of economic plans were drawn up by various groups and individuals, including the National Planning Committee of the Indian National Congress (Visveswarya 1934; Nehru 1946; Banerjee et al. 1944; Thakurdas et al. 1944; Agarwal 1960), that anticipated most post-war debates and ideas on development objectives, strategy and policy in academia and international organizations. The plans saw poverty alleviation as the basic development objective, outlined a 'basic needs' strategy and covered 'redistribution with growth', the development of agriculture versus industry, heavy industry-based industrialization and import substitution, the respective roles of large- and small-scale industries and of the state versus the market (see Srinivasan 2001).

The Rise and Fall of the Planning Syndrome

With the setting up of the Planning Commission in the 1950s India embarked on a public sector dominated by heavy industry and an import-substituting industrialization strategy as the answer to alleviate its ancient poverty. Professor P.C. Mahalanobis (1953, 1955), a distinguished statistician and the father of Indian planning, provided its rationale in a formal model, taken largely from the model that the Soviet economist Fel'dman had developed for Stalin's industrialization strategy. This showed that, with a binding foreign exchange constraint (which, on the basis of the export pessimism generated by the experience of the Great Depression, was assumed to confront India) independent of a savings

constraint to limit the growth rate of the economy, a higher sustainable development path could be attained by using limited foreign exchange to import (and so support the industrial structure vertically) machines to make machines, until India was producing everything she needed, except for the raw materials that could not be obtained domestically (see Bhagwati and Chakravarty 1969; Lal 1972a).

The Perspective Planning Division of the Planning Commission, headed by its intellectually curious and energetic head, Pitamber Pant, and the branch of Mahalanobis' Indian Statistical Institute (ISI) attached to it, then became the centre of intense intellectual debate. In the 1960s it employed a growing number of Indian economists trained in Western universities (Bhagwati, Bardhan, Minhas, Parikh, Srinivasan, Tendulkar among others), and in association with a programme set up by Rosenstein Rodan at Massachusetts Institute of Technology (MIT) became host to a galaxy of foreign economists (Swan, Reddaway, Lewis, Little and Harberger). The Delhi School of Economics, under the leadership of K.N. Raj, engaged Chakravarty and Sen, and at the Finance Ministry I.G. Patel invigorated the newly established Indian Economic Service by engaging V.K. Ramaswami and Manmohan Singh as economic advisors. Meanwhile, the USAID mission was headed by J.P. Lewis, and the number of foreign economists visiting and participating in the economic debates of the time expanded to include Milton Friedman and Peter Bauer.

The Mahalanobis model was to form the analytical basis for India's second Five Year Plan. The Planning Commission had convened a panel of economists to discuss its framework, and most of them endorsed the broad objectives and strategy of the plan. The only dissenting voice was that of B.R. Shenoy, who questioned, amongst other issues, the massive deficit financing on which the plan depended. In this he was supported by two of the visiting foreign economists, Peter Bauer and Milton Friedman. Whilst Komiya (1959) and Bronfrenbrenner (1960) provided explicit critiques of the Mahalanobis model. But most of these criticisms were disregarded by the prevailing intellectual consensus in favour of dirigiste, state-led planning, though the technocratic basis of the planning models on which it was based was increasingly questioned by Indian economists (see Rudra 1975).

With the emergence of what J.P. Lewis (1963) accurately described as a 'quiet crisis' in India, engendered by the foreign exchange crisis caused by the fiscal expansion the dissenters had predicted (which had led to draconian foreign tradecum- exchange and price controls), new voices arose in the 1960s providing the intellectual basis for the subsequent neoclassical resurgence in development economics. Developing ideas presaged in the writings of James Meade and Harry Johnson, two Indian economists, Jagdish Bhagwati (who was at the ISI) and V.K. Ramaswami, economic advisor at the Ministry of Finance, produced a path-breaking paper that began the process of separating the case for free trade from that for laissez-faire (Bhagwati and Ramaswami 1963). In a series of papers with T.N. Srinivasan (also at the ISI), they established the modern theory of trade and welfare which shows that most of the arguments for protection are second best as they depend upon 'domestic distortions' in the working of the price mechanism, which are best dealt with by direct domestic taxes and subsidies rather than the indirect method of protection.

Two major books, by Bhagwati and Desai (1970) and Bhagwati and Srinivasan (1975), written as part of two large-scale multi-country comparative studies of trade and industrialization directed by I.M.D. Little, T. Scitovsky and M. Fg. Scott for the Organization for Economic Cooperation and Development (OECD), and by J. Bhagwati and A. Krueger for the National Bureau of Economic Research (NBER), provided a detailed empirical analysis of the relevance of this newly developed theory, besides documenting the immense inefficiency and corruption that the dirigiste planning system had engendered. This marked the beginning of the end of the planning syndrome that had held Indian economists in thrall for nearly a century. Furthering this disenchantment was the disappointing performance of Indian industry where the net

effect of the control system was shown to be a capital-intensive bias and low or negative growth of total factor productivity in post-Independence industrial performance (I.J. Ahluwalia 1985).

Moreover, Manmohan Singh (1964), in a detailed study of Indian exports, had shown that the export pessimism underlying the assumption of a foreign-exchange constraint in the Mahalanobis model was unjustified, as it was not lack of external demand but the consequences of India's domestic economic policies that had led to the disappointing Indian export performance.

Nor was the panacea offered by the Gandhians – which was promulgated with reservations for various small-scale industries (particularly cotton textiles) on the grounds that they promoted employment growth – found to be valid. Dhar and Lydall (1961) in an empirical study of these industries showed that these smallscale industries were technically inefficient than their larger modern brethren because they used both more labour and capital per unit of output produced.

The planners' belief that the public sector, given monopoly production rights in the 'commanding heights' of the economy, would be dynamic and through rising profits augment domestic savings was discredited. Numerous official empirical studies documented the growing inefficiency of the public sector and its growing drain on the nation's savings. As part of the debate on their reform which came to the fore in the 1970s, two major manuals of project evaluation were developed to improve the efficiency of the public sector. One was produced for the UN's Industrial Development Organization by P. Dasgupta, A.K. Sen and S. Marglin the other for the OECD by I.M.D. Little and J.A. Mirrlees. With the implicit adoption of the latter by a newly set up Project Appraisal Division in the Planning Commission, Lal (1980) produced the first comprehensive set of 'shadow prices' based on the 'world price rule' for use in the evaluation of public projects in India. But the social cost–benefit analysis they were meant to support soon descended into social cosmetic analysis, as politicians continued to choose and run public projects for rent-seeking reasons rather than social profitability. It was not until the fiscal-cum-foreign exchange crisis of 1991 that planning, and the system of controls on industry and foreign trade it had engendered, finally came to a *de facto* if not *de jure* end. The market increasingly came to replace the plan, and a programme of privatization was slowly and fitfully begun.

Transforming Agriculture

An implicit assumption of the Mahalanobis framework was that agriculture could be left alone, merely being a source of 'surplus labour' and of the limited savings and foreign exchange for the heavy industrialization strategy. By the mid-1960s this neglect had led to a severe food crisis. The transformation of agriculture, which until then had been seen largely as a means of promoting equity through land reforms, then became a matter of debate.

Nationalist and Marxist literature in India, basing itself on the perceived outcomes of the laissez-faire period of colonial rule, had maintained that the commercialization of agriculture through the creation, definition and enforcement of saleable and mortgageable land rights, and the integration of the internal economy through the railways had led to an increased concentration of land, the proletarianization of the peasantry and the growth of landless labour and a shift to cash crops from foodgrains, which in turn had led to famine. Subsequent research (summarized in Kumar and Desai 1983, and Lal 1988), has questioned the empirical bases of these beliefs, whilst Sen (1981a) has argued that the periodic famines that have blighted the subcontinent over the millennia were not due to a shortage of food but to 'exchange entitlement failures'. Whenever the monsoon failed there was a drastic fall in the demand for landless labour and thence wages, leading to a reduction in 'exchange entitlement' in terms of food, which in extremity would lead to a famine. The British had already realized this at the end of the nineteenth century, when they set up a famine code whereby, when the rains failed, local District Commissioners were empowered to fund food-for-work public works to provide

the necessary exchange entitlements. As a result, apart from the 1944 famine in Bengal, which was caused by disruptive wartime conditions, India did not see serious famines in the twentieth century.

One of the implicit assumptions underlying the neglect of agriculture in the early plans was that peasants were not subject to economic incentives. Detailed empirical studies by Dharm Narain (1965) and Raj Krishna (1963) of peasant response to the changing relative prices of crops shows that they behaved like *homo economicus* by shifting cropping patterns to crops with higher expected relative prices.

A second tenet (following the famous Arthur Lewis model of a dual economy) was the existence of vast pools of 'surplus labour' in agriculture which could be removed for industrialization without affecting agricultural output. Mehra (1966) provided empirical content by using farm management studies to estimate the surplus labour time available in various states in India. But these and other studies estimating surplus labour did not take account of the wage at which people are willing to work, or the leisure–income choice facing rural workers. They assumed that they would continue to work for an unchanged wage up to a normal number of working hours per day. But, as Sen (1966) showed, even in an overpopulated country, 'surplus labour' – in the sense of a perfectly elastic supply of labour at a constant wage – would imply that leisure was an inferior good. Empirical studies estimating wage elasticities for rural labour in India soon showed that this assumption was invalid (Bardhan 1979, 1984a; Binswanger and Rosenzweig 1984; Lal 1989).

The means to transform Indian agriculture have not changed since the 1893 report by J. Volcker (1893), consultant chemist to the Royal Agricultural Society. His remedies were: irrigation, fertilizers, better seeds and improvements in land tenure. This has been the conventional wisdom on raising Indian agricultural productivity ever since.

An empirical finding from the Indian farm management studies that there was an inverse relationship between the size of farm and productivity per hectare (Sen 1975, Appendix C) was used to argue for land reforms that would break up large farms and create small, family-labour based and family-owned peasant farms, which would promote both equity and efficiency (Rudra and Sen 1980). However, Bhalla and Roy (1988) showed that, once appropriate adjustments were made for differences in land quality, the inverse relationship between farm size and productivity disappears. This undermined the case for land reform in India.

Lal (1988, 2005, 2006) argued that the Malthusian view that population pressure would lead to a stagnation of rural and industrial wages was invalid, as the alternative Boserupian perspective (Boserup 1965) provided a better description of the changing fortunes of Indian agriculture. Boserup argued that population pressure both induces and facilitates the adoption of more intensive forms of agriculture. She identifies the differing input-per-hectare requirements of different agrarian systems by the frequency with which a particular piece of land is cropped. Thus settled agriculture is more labour- and capital- intensive than nomadic pastoralism, which is in turn more intensive in these inputs than hunting and gathering or the slash-andburn agriculture practised until recently in parts of Africa and the tribal regions of India. Contrary to Malthusian presumptions, population growth leads to the adoption of more advanced techniques that raise yield per acre. Because these new techniques require increased labour effort, they will not be adopted until rising population reduces the per capita food output that can be produced with existing techniques and forces a change. Lal marshals empirical evidence to show that Indian agriculture's long trajectory fits this Boserupian framework, with the population expansion beginning from the early 1900s leading in the post-Independence period to an intensification of agriculture, and with the availability of the new high-yielding varieties (HYV) of seeds, to the Green Revolution in the late 1960s and 1970s.

Many of those adhering to the Marxist canon believed and hoped that the bulk of the income

gains arising from the massive increases in output brought about by the Green Revolution would accrue to landowners, and that rural real wages would stagnate, leading to the revolution turning red. But the evidence showed that with the massive shift in the labour-demand curve that resulted from the new technology there was a marked rise in rural real wages (Ahluwalia 1978; Lal 1976, 1989).

As the new HYV technology required an assured water supply along with high dosages of fertilizers, Volcker's other major means of transforming Indian agriculture, namely irrigation, came to the fore. Surface irrigation was expanded during the Raj (the period of British rule in India), particularly in the drier regions where the marginal social returns from irrigation were likely to be the highest. But these schemes were devised by engineers and their direct and indirect economic effects were not estimated, leading in many cases to long-term losses through salination, waterlogging and the creation of malarial swamps (see Whitcombe 1971). In the 1970s two studies of irrigation – of a major surface water scheme, the Bhakra dam, by Minhas et al. (1972) and of groundwater (well) irrigation in the Deccan plateau by Lal (1972b) – provided economic analyses of irrigation and their optimal design.

One of the deleterious effects of the system of protection set up during the Permit Raj was the heavy implicit tax on agriculture. From 1965 efforts were made to correct this by price supports to farmers, which led to an improvement in the terms of trade. But this changed again in the 1980s with growing but inefficient input subsidies becoming the main form of supporting agriculture. With the post- 1991 liberalization of trade largely affecting industrial products, part of the bias against agriculture was removed. The debate then moved to removing the remaining agricultural protection (particularly for cereals), with proponents (Gulati 1998) arguing for domestic prices of agricultural products to be aligned with world prices to allow agriculture to develop in line with its revealed comparative advantage, and opponents (Patnaik 1996) arguing against, on grounds of food security.

Poverty and Income Distribution

A continuing debate concerns the effects on income distribution and poverty of rapid capitalist growth. Indian economists have been in the forefront in both setting out the conceptual basis as well as the measurement of poverty (see Sen 1976, 1981a, b; Dandekar and Rath 1971; Bardhan and Srinivasan 1974; Srinivasan 1983). The internationally adopted headcount ratio (HCR) of the poor below a nutritionally based poverty line of 15 rupees per capita (at 1960–1961 prices) was based on this efflorescence of research in the 1970s (but see Sukhatme 1978; Srinivasan and Bardhan 1988). The continuing debate has centred on whether rapid (capitalist) growth would alleviate poverty without adverse effects on income distribution, or whether more direct methods of redistribution would be needed to alleviate poverty and prevent any worsening of income distribution. A summary of the evidence from these numerous studies based on two large national surveys undertaken by the official National Sample Survey and those undertaken by the unofficial National Council of Applied Economic Research (NCAER) is provided in Lal et al. (2001a, b). There seems to be no clear trend in the Gini coefficient during the 50 years since Independence in 1947, whilst the fluctuating HCR for poverty shows no marked change until the acceleration of the growth rate after the economic liberalization of the 1990s, since when there has been a fall of varying magnitudes, depending upon which study one trusts.

The nationalist-cum-Marxist School unsurprisingly has argued that 'trickle down' would not alleviate poverty. Given the abysmally poor growth record during the planning period, which was characterized as the Hindu rate of growth (of about 1.5% a year in per capita income from the 1950s to early 1980s) it would six have been surprising if there had been any marked alleviation of India's mass structural poverty. Nevertheless, influential voices on the Left articulated a critique of the capitalist growth process. This critique, purportedly supported by Indian data, was soon shown to be false. Thus it was argued that the

alleviation of poverty and equitable growth within the 'existing institutional framework' would not occur because of an increased concentration of land (Raj 1976; refuted by Sanyal 1977a, b); the increasing proletarianization of the countryside (Raj 1976; refuted by Visaria 1977); increasing rural indebtedness and usury (disputed by Ghatak 1976); a continual improvement in the agricultural terms of trade which damaged industrial development (Bagchi 1970; Chakravarty 1974; Sau 1981; Vaidyanathan 1977; and Mitra 1977), which were critiqued by Desai (1981); and the inimical effects of foreign investment (Sau 1981) which is countered in Lal et al. (1975). These are now seen as shibboleths, particularly after the death of the countries of 'really existing socialism' and the economic liberalizations of the 1990s. The intemperate debate this provoked between the left-wing radicals and neoclassical liberalizers showed up the ideological nature of this debate, with Rudra (1991) stating: 'I put my ideological cards on the table. I hate capitalism', and Srinivasan (1992) rightly responding: 'In Rudra's value system competition, without which the market economy cannot efficiently function, is an instrument with a negative value connotation. In this he would be in the good company of monopolists and oligopolists and state capitalists of the world who would also dearly love to eliminate competition!'

While growth is being increasingly accepted as necessary for the sustainable alleviation of mass structural poverty (see Tendulkar 1998), Lal and Myint (1996) argue that two other forms of poverty, destitution and conjunctural poverty, require income transfers, though not necessarily public ones. Though Dasgupta (1993) claims to be about destitution, it is more about mass structural poverty and income distribution (Srinivasan 1994). The only study of destitution (Lipton 1983) based on village studies found no obvious correlates to identify an extremely heterogenous group. Thus Dasgupta's reasonable assertion that widows become destitute was belied by the evidence in Drèze and Srinivasan (1995).

Public policy has thus sought to deal with the third triad of poverty, conjunctural poverty, which is largely associated with climatic variations through a continuation of the Raj's famine code to prevent famine and by rural employment guarantee schemes to offset seasonal unemployment by offering jobs on public works at a wage only the needy will accept, which because of self-targeting have been shown to be efficacious (Ravallion 1991).

The major advocate of the direct route for poverty alleviation (where the three categories distinguished above are amalgamated) remains Sen (1981a, b), whose earlier empirical evidence on the superiority of this route in low-growth economies (Sri Lanka) and regions (Kerala in India) was questioned by Bhalla and Glewwe (1986). The debates in Drèze and Sen (1989) concentrate on the public provision of food for the malnourished and the merit goods of health and education. But empirical studies of the nearly 50-year-old public programmes to deal with these aspects do not provide much hope for success (Parikh 1993; World Bank 2000; PROBE Team 1999). Similarly, the dismal state of publicly owned and operated infrastructure (Ahluwalia 1998; Ahluwalia and Little 1998) has led to a search for decentralized private solutions to provide these 'public goods' with public funding (Mitra 2006; Bardhan and Mookherjee 2006).

Political Economy and Institutions

With the growing corruption engendered by the Permit Raj, there have been attempts to measure what Krueger (1974) has designated as the 'rent-seeking society'. Her attempts at measuring the rents created by the Permit Raj in India has been supplemented by other studies (see Acharya 1985; Mohammad and Whalley 1984), whilst her rent-seeking model has been expanded by Bhagwati and Srinivasan to encompass a whole host of what they term 'directly unproductive activities' (Bhagwati and Srinivasan 1980).

A large political economy literature has arisen to explain the economic outcomes in India's democratic polity. Much of this has a Marxist lineage (Raj 1973; Jha 1980; Bardhan 1984b). Lal (1984, 1988, 2005) on the other hand has developed a model of 'the predatory state' which maximizes

net revenue and has argued that the successive empires in north India were predatory states that fell when they attempted to extract more than the natural 'rent' the economic system could provide. Lal (1987) and Lal and Myint (1996) also provide a theory which seeks to explain the role of crises in generating economic reforms in previously repressed economies. This is borne out by the liberalization undertaken in the face of a serious fiscal, foreign exchange and inflationary crisis in 1991 caused by the cumulative effects of the dirigisme of the Permit Raj.

There have also been attempts to explain various institutions that have shaped economic outcomes: the caste system (Lal 1988, 2005) as a means of tying scarce labour down to abundant land, and a theory of interrelated factor markets which seeks to explain seemingly inefficient institutions like sharecropping, attached labour, and usurious interest rates as second-best adaptations to problems of risk and the uncertainty to which tropical agriculture is subject (Bardhan 1980; Bardhan and Rudra 1978; Srinivasan et al. 1997; Basu 1983).

The Macroeconomy

Post-Independence India followed an orthodox monetary policy based on the system of fiscal and monetary accounting left by the Raj. In the 1980s, however, in order to push up the growth rate it began to undertake risky macroeconomic policies, and, with the crisis of 1991, macroeconomic issues came to the fore. The best account of India's macroeconomy since Independence was provided by Joshi and Little (1994), whilst Bhagwati and Srinivasan (1993) and Virmani (2001) provide analyses of the genesis of the crises and the lineaments of the partial and still incomplete economic liberalization that occurred in the wake of the crisis.

With the opening of the economy and (by the standards of the planning era) large inflows of foreign capital, India faced the prospect of Dutch disease – with a rise in the real exchange rate reducing the profitability of tradable relative to nontraded goods. The authorities responded by sterilizing these inflows and building up large foreign-exchange reserves, thus stalling an appreciation of the nominal exchange rate, to maintain the competitiveness of Indian exports (which, after their post-Independence stagnation, in the 1990s began to take off with the gradual integration of India into the world economy). Because of the continuing large fiscal deficits, particularly of the states in the Indian federation (Lal et al. 2001a, b), the government was also reluctant to open the capital account for fear of these deficits spilling over and causing another foreign debt crisis. A lively debate began in the early part of the twenty-first century on the correct monetary and exchange-rate policy for India to follow in the light of the continuing build-up in foreign exchange reserves. Lal et al. (2003) argued for liberalizing the capital account and floating the rupee. Joshi and S. Sanyal (2004) demurred, arguing for capital account controls and a managed exchange rate, largely on grounds of exchange-rate protection. The debate is still ongoing as of 2007, and the government has reconstituted an official committee which in the late 1990s had cautioned on opening the capital account.

The economic debates in India have thus moved on to what are no longer distinctively Indian issues, and local contributions are now less likely to be groundbreaking or to deal uniquely with issues in the current debates on development in the subcontinent.

See Also

- ▶ Agriculture and Economic Development
- ▶ Development Economics
- ▶ Planning
- ▶ Poverty
- ▶ Poverty Alleviation Programmes

Bibliography

Acharya, S. 1985. *Report on the underground economy*. New Delhi: National Institute of Public Finance and Policy.

Agarwal, S.N. 1960. *Principles of Gandhian planning*. Bombay: Kitab Mahal.

Ahluwalia, M.S. 1978. Rural poverty and agricultural performance in India. *Journal of Development Studies* 14: 298–323.

Ahluwalia, I.J. 1985. *Industrial growth in India: Stagnation since the mid-1960s*. New Delhi: Oxford University Press.

Ahluwalia, M.S. 1998. Infrastructure development in India's reforms. In Ahluwalia and Little (1998).

Ahluwalia, I.J., and I.M.D. Little (eds.). 1998. *India's economic reforms and development: Essays for Manmohan Singh*. New Delhi: Oxford University Press.

Bagchi, A.K. 1970. Long term constraints on India's industrial growth 1951–1968. In *Economic development in South Asia*, ed. E.A.G. Robinson and M. Kidron. London: Macmillan.

Banerjee, B.N., G.D. Parikh, and V.M. Tarkunde. 1944. *People's plan for economic development of India*. Bombay: Indian Federation of Labour.

Bardhan, P. 1979. Labor supply functions in a poor agrarian economy. *American Economic Review* 69: 73–83.

Bardhan, P. 1980. Interlocking factor markets and Agrarian development: A review of the issues. *Oxford Economic Papers* 32: 82–98.

Bardhan, P. 1984a. *Land, labour and rural poverty*. New Delhi: Oxford University Press.

Bardhan, P. 1984b. *The political economy of development in India*. Oxford: Blackwell.

Bardhan, P., and D. Mookherjee. 2006. Decentralization and accountability in infrastructure delivery in developing countries. *Economic Journal* 116: 107–133.

Bardhan, P., and A. Rudra. 1978. Interlinkage of land, labour, and credit relations: An analysis of village survey data in East India. *Economic and Political Weekly* 13: 367–384.

Bardhan, P., and T.N. Srinivasan (eds.). 1974. *Poverty and income distribution in India*. Calcutta: Statistical Publishing Society.

Basu, K. 1983. The emergence of isolation and interlinkage in rural markets. *Oxford Economic Papers* 35: 262–280.

Bhagwati, J., and S. Chakravarty. 1969. Contributions to Indian economic analysis: A survey. *American Economic Review* 59: 2–73.

Bhagwati, J., and P. Desai. 1970. *India: Planning for industrialization*. London: Oxford University Press.

Bhagwati, J., and V.K. Ramaswami. 1963. Domestic distortions, tariffs, and the theory of optimum subsidy. *Journal of Political Economy* 71: 44–50.

Bhagwati, J., and T.N. Srinivasan. 1975. *Foreign trade regimes and economic development: India*. New York: Columbia University Press.

Bhagwati, J., and T.N. Srinivasan. 1980. Revenue seeking: A generalization of the theory of tariffs. *Journal of Political Economy* 88: 1069–1087.

Bhagwati, J., and T.N. Srinivasan. 1993. *India's economic reforms*. New Delhi: Ministry of Finance, Department of Economic Affairs.

Bhalla, S., and P. Glewwe. 1986. Growth and equity in developing countries: A reinterpretation of the Sri Lanka experience. *World Bank Economic Review* 1: 35–63.

Bhalla, S., and P. Roy. 1988. Misspecification in farm productivity analysis: The role of land quality. *Oxford Economic Papers* 40: 55–73.

Binswanger, H., and M. Rosenzweig (eds.). 1984. *Contractual arrangements, employment and wages in rural labor markets in Asia*. New Haven: Yale University Press.

Boserup, E. 1965. *The conditions of agricultural growth*. London: Allen & Unwin.

Bronfrenbrenner, M. 1960. A simplified Mahalanobis development model. *Economic Development and Cultural Change* 9: 45–51.

Chakravarty, S. 1974. Reflections on the growth process in the Indian economy. In *Some problems of indian economic policy*, 2nd ed, ed. C.D. Wadhwa. New Delhi: Tata-McGraw Hill.

Dandekar, V.M., and N. Rath. 1971. *Poverty in India*. Poona: Indian School of Political Economy.

Dasgupta, P. 1993. *An inquiry into well-being and destitution*. Oxford: Clarendon.

Desai, A.V. 1981. Factors underlying the slow growth of Indian industry. *Economic and Political Weekly* 16(10–12): 381–392.

Dhar, P.N., and H.F. Lydall. 1961. *The role of small enterprises in Indian economic development*. Bombay: Asia Publishing House.

Drèze, J. and P.V. Srinivasan. 1995. *Widowhood and poverty in India: Some inferences from survey data*. Discussion Paper No. 62. Development Economics Research Program, London School of Economics.

Drèze, J., and A.K. Sen. 1989. *Hunger and public action*. Oxford: Clarendon.

Dutt, R.C. 1904. *The economic history of India*, vol. 2. New Delhi: Publications Division, Ministry of Information and Broadcasting.

Ghatak, S. 1976. *Rural money markets in India*. New Delhi: Macmillan.

Gulati, A. 1998. Indian agriculture in an open economy: Will it prosper? In Ahluwalia and Little (1998).

Jha, P.S. 1980. *The political economy of stagnation*. New Delhi: Oxford University Press.

Joshi, V., and I.M.D. Little. 1994. *India: Macroeconomics and political economy 1964–1991*. Washington, DC: World Bank.

Joshi, V. and S. Sanyal. 2004. Foreign inflows and macroeconomic policy in India. In *NCAER/Brookings: India Policy Forum 2004*, Washington, DC: Brookings Institution.

Komiya, R. 1959. A note on Professor Mahalanobis' model of Indian economic planning. *Review of Economics and Statistics* 41: 29–35.

Krishna, R. 1963. Farm supply response in India: Pakistan. *Economic Journal* 73: 477–487.
Krueger, A.O. 1974. The political economy of the rent-seeking society. *American Economic Review* 64: 291–303.
Kumar, D. and M. Desai. eds. 1983. *The Cambridge economic history of India*, vol. 2: c. 1750–1970. Cambridge, UK: Cambridge University Press.
Lal, D. 1972a. The foreign exchange bottleneck revisited. *Economic Development and Cultural Change* 20: 722–730.
Lal, D. 1972b. *Wells and welfare: An exploratory cost-benefit study of the economics of small scale irrigation in Maharashtra*. Paris: OECD.
Lal, D. 1976. Agricultural growth, real wages and the rural poor in India. *Economic and Political Weekly* 11(26): A47–A61.
Lal, D. 1980. *Prices for planning: Towards the reform of Indian planning*. London: Heinemann Educational Books.
Lal, D. 1984. *The political economy of the predatory state*, Discussion Paper DRD 105, Development Research Department, Washington, DC: World Bank.
Lal, D. 1987. The political economy of economic liberalization. *World Bank Economic Review* 1: 273–299.
Lal, D. 1988. *The Hindu equilibrium, vol. 1: Cultural stability and economic stagnation, India c. 1500 BC–AD 1980*. Oxford: Clarendon.
Lal, D. 1989. *The hindu equilibrium, vol. 2: Aspects of Indian labour*. Oxford: Clarendon.
Lal, D. 2005. *The Hindu equilibrium, India c. 1500 BC–2000 AD*, abridged and revised edn. Oxford: Oxford University Press.
Lal, D. 2006. India: Population change and its consequences. *Population and Development Review* 32-(Supplement), 145–182.
Lal, D., and H. Myint. 1996. *The political economy of poverty, equity and growth: A comparative study*. Oxford: Clarendon.
Lal, D., M. Cave, P. Hare, and J. Thompson. 1975. *Appraising foreign investment in developing countries*. London: Heinemann Educational Books.
Lal, D., S. Bhide, and D. Vasudevan. 2001a. Financial exuberance: Savings deposits, fiscal deficits and interests rates in India. *Economic and Political Weekly* 37: 4196–4203.
Lal, D., R. Mohan, and I. Natarajan. 2001b. Economic reforms and poverty alleviation: A tale of two surveys. *Economic and Political Weekly* 36: 1017–1028.
Lal, D., S. Bery, and D.K. Pant. 2003. The real exchange rate, fiscal deficits and capital flows, India: 1981–2000. *Economic and Political Weekly* 38: 4965–4976.
Lewis, J.P. 1963. *Quiet crisis in India*. Bombay: Asia Publishing House.
Lipton, M. 1983. *Labour and poverty*. World Bank Staff Working Papers No. 616, Washington, DC: World Bank.
Mahalanobis, P.C. 1953. Some observations on the process of growth in national income. *Sankhya* 12: 307–312.
Mahalanobis, P.C. 1955. The approach of operational research to planning in India. *Sankhya* 16: 3–130.
Mehra, S. 1966. Surplus labour in Indian agriculture. *Indian Economic Review* 1(1): 111–129.
Minhas, B.S., K. Parikh, and T.N. Srinivasan. 1972. *Scheduling the operations of the Bhakra system*. Calcutta: Statistical Publishing Society.
Mitra, A. 1977. *Terms of trade and class relations*. London: Frank Cass.
Mitra, B.S. 2006. Grassroots capitalism thrives in India. In *2006 index of economic freedom*, ed. M. Miles, K.R. Holmes, and M.A. O'Grady. Washington, DC: Heritage Foundation.
Mohammad, J., and J. Whalley. 1984. Rent seeking in India: Its costs and policy significance. *Kyklos* 37: 387–413.
Naoroji, D. 1901. *Poverty and Unbritish rule in India*. London: Swan Sonnenschein and Co.
Narain, D. 1965. *The impact of price movements on areas of selected crops in India 1900–39*. Cambridge, UK: Cambridge University Press.
Nehru, J. 1946. *The discovery of India*. New York: John Day Co.
Parikh, K.S. 1993. *Who gets how much from the PDS – How effectively does it reach the poor?* Mumbai: Indira Gandhi Institute of Development Research.
Patnaik, U. 1996. Export oriented agriculture and food security in developing countries and India. *Economic and Political Weekly* 31: 2429–2449.
PROBE Team. 1999. *Public report on basic education in India*. New Delhi: Oxford University Press.
Raj, K.N. 1973. The politics and economics of intermediate regimes. *Economic and Political Weekly* 8: 1189–1198.
Raj, K.N. 1976. Trends in rural unemployment in India. *Economic and Political Weekly* 11: 1781–1862.
Ravallion, M. 1991. Reaching the poor through public employment: Arguments, evidence and lessons from South Asia. *World Bank Research Observer* 6(2): 153–175.
Rudra, A. 1975. *Indian plan models*. New Delhi: Allied Publishers.
Rudra, A. 1991. Privatization and deregulation. *Economic and Political Weekly* 27: 2933–2936.
Rudra, A., and A.K. Sen. 1980. Farm size and labour use: Analysis and policy. *Economic and Political Weekly* 15: 391–394.
Sanyal, S.K. 1977a. Trends in rural unemployment: A comment. *Economic and Political Weekly* 12: 145–148.
Sanyal, S.K. 1977b. Trends in some characteristics of landholdings: an analysis for a few states I and II. *Sarvekshana* 1(1–2).
Sau, R. 1981. *India's economic development: Aspects of class relations*. New Delhi: People's Publishing House.

Sen, A.K. 1966. Peasants and Dualism, with and without surplus labour. *Journal of Political Economy* 74: 425–450.
Sen, A.K. 1975. *Employment, technology and development*. Oxford: Oxford University Press.
Sen, A.K. 1976. Poverty: An ordinal approach to measurement. *Econometrica* 44: 219–231.
Sen, A.K. 1981a. *Poverty and famines: An essay on entitlement and deprivation*. Oxford: Clarendon.
Sen, A.K. 1981b. Public action and the quality of life in developing countries. *Oxford Bulletin of Economics and Statistics* 43: 287–318.
Singh, M. 1964. *India's export trends and the prospects for self contained growth*. Oxford: Clarendon.
Sivasubramonian, S. 2000. *The national income of India in the twentieth century*. New Delhi: Oxford University Press.
Srinivasan, T.N. 1983. Hunger: defining it, estimating its global incidence and alleviating it. In *Role of markets in the world food economy*, ed. D. Gale Johnson and E.G. Schuh. Boulder: Westview Press.
Srinivasan, T.N. 1992. Privatization and deregulation. *Economic and Political Weekly* 27: 843–848.
Srinivasan, T.N. 1994. Destitution: A discourse. *Journal of Economic Literature* 32: 1842–1855.
Srinivasan, T.N. 2001. *economic policy and state intervention*. New Delhi: Oxford University Press.
Srinivasan, T.N., and P. Bardhan (eds.). 1988. *Rural poverty in south Asia*. New York: Columbia University Press.
Srinivasan, T.N., C. Bell, and C. Udry. 1997. Rationing, spillover, and interlinking in credit markets: The case of rural Punjab. *Oxford Economic Papers* 49: 557–585.
Sukhatme, P.V. 1978. Assessment of adequacy of diets of different economic levels. *Economic and Political Weekly* 13(Special number), 1373–1384.
Tendulkar, S. 1998. Indian economic policy reforms and poverty: An assessment. In *Ahluwalia and Little* (1998).
Thakurdas, P.J., G. Birla, A. Dalal, S. Ram, K. Lalabhai, A. Shroff, and J. Matthai. 1944. *A plan of economic development of India*. London: Penguin.
Vaidyanathan, A. 1977. Performance and prospects of crop production in India. *Economic and Political Weekly* 2: 1355–1368.
Virmani, A. 2001. *India's 1990–91 crisis: Reforms, myths and paradoxes*. New Delhi: Planning Commission, Government of India.
Visaria, P. 1977. Trends in rural unemployment in India: A comment. *Economic and Political Weekly* 12: 145–148.
Visveswarya, M. 1934. *Planned economy for India*. Bangalore: Bangalore Press.
Volcker, J.A. 1893. *Report on the development of Indian agriculture*. Bombay.
Whitcombe, E. 1971. *Agrarian conditions in Northern India*. Berkeley and Los Angeles: University of California Press.
World Bank. 2000. *India: Reducing poverty, accelerating development*. New Delhi: Oxford University Press.

Indian Economic Development

Arvind Panagariya

Abstract

Four or arguably five phases can be identified in India's post-independence economic experience. The first phase in which institutions were put in place and policies were relatively liberal saw moderate growth, but this was stifled by command and control policies in the second phase. More recent liberalization has seen renewed increased in the growth levels, and it is argued that this should continue beyond the 2008–9 economic crisis. However manufacturing, especially labour-intensive sectors, continue to grow slowly, growth is heavily reliant on the service sector, and a disproportionately large workforce remains engaged in inefficient agricultural production.

Keywords

Command economy; India; Liberalization

JEL Classifications

O53

Post-independence India is one of the most fascinating case studies in economic development. During the six decades since independence in 1947, it has experimented with a diverse set of economic policies within a parliamentary democracy and an institutional framework that has remained unchanged except in details to accommodate the policy changes undertaken. Panagariya (2008), on which this article draws heavily, offers a comprehensive discussion of the twists and turns of the policy and the accompanying ups and downs in the economy.

Four Phases of Growth

India became independent in 1947 and formally launched its economic development programme

in the financial year 1951–2. (Data on India usually relate to its fiscal year, which begins on 1 April and ends on 31 March. Therefore, 1951–2 refers to the period from 1 April 1951 to 31 March 1952.) From the beginning, the overarching objective of the government was the eradication of poverty. Rapid growth was seen as a means to achieving that objective, although it was sometimes stated as an objective in itself in view of the close link between it and poverty alleviation, which Indian analysts in the government saw at an early stage.

The overall economic performance and its link to policies during the six decades are best explained by dividing the period between 1951–52 and 2007–08 into four distinct phases. (Throughout, a period such as 1951–65 refers to the years from 1951–2 to 1964–5 with end-point years included.)

- Phase I (1951–65) with an average annual growth rate of 4.1 per cent.
- Phase II (1965–81) with an average annual growth rate of 3.2 per cent.
- Phase III (1981–8) with an average annual growth rate of 4.6 per cent.
- Phase IV (1988–2008) with an average annual growth rate of 6.6 per cent.

Phase IV can be further subdivided into 15 years spanning 1988–2003 and five years spanning 2003–08, with average annual growth rates of 5.8 and 8.8 per cent respectively.

Phase I (1951–65): Take-Off Under a Liberal Regime

The preservation of India's independence was one of the foremost goals of Pundit Jawaharlal Nehru, the first prime minister of India. He also saw economic independence from the world markets as essential for preserving political independence. While this thinking did not necessarily imply a protectionist import policy, it did required progressive realignment of the production basket with the consumption basket so as to eliminate the need for trade. In Nehru's own words, 'The objective for the country as a whole was the attainment, as far as possible, of national self-sufficiency. International trade was certainly not excluded, but we were anxious to avoid being drawn into the whirlpool of economic imperialism' (1946, p. 403). Nehru reasoned that since private entrepreneurs lacked resources to invest in machinery, metals and other heavy industry, the public sector had to play an active role in these sectors. In addition, he considered it necessary to direct larger privatesector enterprises through investment licensing towards sectors of greater social value than just private profitability. The experience of the Soviet Union, considered a success at the time, also led Nehru to adopt a similar system of planning, with the First Five Year Plan launched in 1951–2.

Although the public sector entered manufacturing activity as a major player in heavy industry, and investment licensing was put in place for enterprises investing 10 million rupees or more, the policy regime remained relatively liberal during the 1950s. Applicants obtained licences with relative ease, with few of them complaining during this period. Import licensing had existed since the Second World War but imports were permitted relatively freely, with significant quantities of foreign consumer goods entering the country. Nehru resisted the demands by the left parties for the nationalization of foreign companies, and maintained a liberal foreign investment policy throughout his rule.

This period also saw the major institutions of the country put in place or revamped. A democratic constitution with parliamentary form of government came into force. Bureaucracy took shape, with officers and employees placed at various levels of administration. The police force was expanded. Schools, colleges and universities multiplied, with the government itself becoming a major employer of scientists and researchers.

A major turning point during this period came in 1958 when, reacting to a foreign exchange shortage, the Finance Ministry adopted centralized foreign exchange budgeting. Under this system the ministry estimated the available foreign exchange for each forthcoming six-month period and allocated it administratively across various claimants. This single policy change considerably tightened not only the import policy but also investment licensing: unless foreign exchange was available for imports of machinery and raw material, a

licence could not be issued. By the mid-1960s, the impact of tightening came to be widely recognized, as was reflected in the large number of government committees that were set up to suggest ways in which the licensing procedures could be improved to eliminate the delays.

With relatively free foreign investment and import policy during the 1950s, an expansionary fiscal policy in the early to mid-1960s, rising savings rates, increased population growth and an overall policy regime geared to the national economic interest, India was able to accelerate its growth rate from less than one per cent during the first half of the 20th century to 4.1 per cent during this phase. This was a source of some satisfaction. Nevertheless, as Maddison (1971) rightly notes, when we consider that much of the rest of the world had grown more rapidly during this period, 'India's post-war performance is well below the average for the developing countries.' After systematic examination, Maddison concluded that India had performed below its potential.

Phase II (1965–81): Socialism Triumphs

Nehru died in 1964 and was succeeded by Lal Bahadur Shastri as prime minister. Shastri did not share Nehru's enthusiasm for heavy industry and was keener on agriculture. Although he passed away within 19 months of assuming the reins, he laid down the foundation of the Green Revolution, perhaps the most important achievement of Phase II.

Indira Gandhi, Nehru's daughter, succeeded Shastri. Political compulsions led her to adopt policies that were highly detrimental to growth and poverty alleviation. Her major policy initiatives included:

- nationalization of the major banks, insurance companies, oil companies and mines
- reservation of the most labour-intensive products for exclusive production by small-scale enterprises (SSE), defined as enterprises with approximately $100,000 or less in assets
- a ban on large firms and business houses, defined as entities with approximately $27 million in investment, investing outside a list of 19 core industries, all of them highly capital intensive
- a 40 per cent ceiling on foreign investment in any firm, with a small number of exceptions
- expansion of price and distribution controls which had been introduced in Phase I
- progressively tighter control on imports through licensing
- a virtual ban on the termination of workers under any circumstances in firms with 300 (later revised to 100) or more workers
- a virtual ban on the acquisition and retention of urban land beyond a tight ceiling varying from 500 square metres (in major cities) to 2,000 square metres.

In effect, these command and control policies, reinforced by a series of external shocks (two wars with Pakistan, two episodes of back-to-back droughts and two oil price shocks) resulted in the growth rate plummeting to an average 3.2 per cent in Phase II, from 4.1 per cent in Phase I. Some observers like to attribute the decline in the growth rate entirely to external shocks, but the importance of policies cannot be underestimated. For one thing, the world economy had grown rapidly from 1965 to 1975, when industrial growth in India fell to just 3.3 per cent from more than six per cent from 1951 to 1964. More importantly, the economy of the South Korea, which adopted an aggressively outward-oriented policy regime beginning in the early 1960s and had none of the command and control machinery of India, shot up like a meteor. It annually grew by 9.5 per cent from1963 to 1973 and by 7.2 per cent from 1974 to 1982.

Phase III (1981–8): Liberalization by Stealth

By 1975, Mrs Gandhi had pushed the command and control policies as far as she could. With industrial growth plummeting and industrialists complaining about unused capacity because of the unavailability of raw materials, or unsatisfied demand because of an inability to expand beyond licensed capacity, pressures began to build up to backtrack. Because no one was willing openly to admit that the system had gone too far, the response was gradual, quiet and within the existing policy framework as if by stealth.

The piecemeal liberalization took place in three phases spanning 1975–9, 1979–84 and 1985–9, with each successive phase being more significant than the preceding one. Measures in the first two phases involved some liberalization of imports by exempting selected products from licensing requirements, allowance for capacity expansion, investment delicensing of selected industries, an increase in the investment level below which a licence was not required, expansion of the list of products in which large firms and big business houses were allowed to invest, and broad-bending of licensed capacity, whereby existing capacity could be used to produce products related to those initially authorized. The last phase, implemented under prime minister Rajiv Gandhi, who succeeded his mother after the latter's Sikh guards had assassinated her in October 1984, went further. In addition to measures similar to those just listed, it included significant measures to promote exports, substantial depreciation of the rupee, tax reform and an end to price and distribution controls on a selected set of important commodities. These reforms, accompanied by an expansionary fiscal policy, returned India to more or less the growth rate it had achieved in Phase I.

Phase IV (1988–2008): The Triumph of Reforms

The last three years of the 1980s saw the growth rate accelerate to 7.2 per cent. This acceleration was achieved partly through expansionary fiscal policy. Fiscal deficits, foreign borrowing, external debt-to-GDP ratio and debt–service ratio (interest and principal payments on external debt as a proportion of export earnings) particularly deteriorated in the second half of the 1980s. For example, the debt–service ratio shot up from 18 per cent in 1984–5 to 27 per cent in 1989–90. The hike in the oil price in the wake of the first Iraq war administered the final blow to a deteriorating balance of payments situation. A balance of payments crisis ensued, paving the way for systematic and systemic reforms this time around.

A Tamil terrorist assassinated Rajiv Gandhi while he was campaigning for the 1991 parliamentary elections. This brought prime minister Narasimha Rao to the helm. Taking advantage of the crisis, Rao decided to set India's house in order. He appointed a technocrat, Dr Manmohan Singh, as his finance minister and provided him with the necessary political support to carry out systematic reforms. In one stroke, the new government abolished import licensing on capital goods and raw materials, ended investment licensing on all but a handful of products and initiated the process of opening the country to foreign investors. The highest industrial tariff rate, which was 355 per cent in 1990–1, was steadily brought down, reaching 50 per cent in 1995–6. The government also took steps to reduce the degree of financial repression, open telecommunications to the private sector and grant entry to private carriers in the airline industry.

The Rao government undertook its most significant reforms in the first three years of its tenure. After that, the reforms became piecemeal once again until the National Democratic Alliance (NDA) was given a clear mandate for five years in 1999. Rao lost his mandate in the 1996 elections and was followed by three fragile coalition governments in as many years.

Finally in 1999, under its determined leader Atal Bihari Vajpayee as prime minister, the NDA government returned to systematic reforms. This second wave of reforms, like the first one during the first three years of the Rao government, touched virtually all sectors of the economy except perhaps labour markets. Trade liberalization was accelerated, doors to foreign investors were opened wider in almost all sectors, genuine privatization of public sector enterprises was introduced, interest rates were liberalized, the insurance sector was opened to the private sector with foreign participation permitted, a key reform of the electricity system was introduced, and above all, a major reform of the telecommunications sector through the New Telecom Policy (1999) revolutionized the communications landscape of India. In my view, while the Rao reforms placed India firmly on the six per cent growth path, the Vajpayee reforms paved the way to the current eight to nine per cent growth.

In a surprise result the NDA government lost the 2004 election, paving the way for the current United Progressive Alliance (UPA) led by the Congress Party. Although Dr Manmohan Singh came to the helm as prime minister raising hopes for continued reforms, internal tensions and opposition from the left parties, which provided critical balancing votes for the survival of the government, held the government's hand back. Indeed, sadly, the government has performed quite poorly, failing to implement policies effectively even in areas of agreement. For instance, from the beginning, the UPA had singled out infrastructure building as its top priority. Yet it ended up considerably slowing down the progress in such critical areas as road building and electricity, where the NDA government had built up substantial momentum. And of course in the critical area of labour regulation, the government pre-committed itself to not undertaking reforms. Even in international trade, a largely non-controversial area, liberalization has come to a standstill.

Has India Moved into Phase V?

A plausible case can be made that starting in 2003–04, India has entered a new phase. Growth during the five-year period spanning 2003–08 averaged 8.8 per cent. This is a full three percentage points higher than the 5.8 per cent rate achieved from 1988–89 to 2002–03. As we shall see below, the economic transformation during the last five years has been unprecedented.

Sceptics argue that the current acceleration is a temporary aberration from the steady-state growth of six per cent. The likely decline in the growth rate in the financial year 2008–09, almost entirely as a result of the global economic crisis, has strengthened this argument. Yet my own view is that over the longer run, say the next ten to 15 years, India will sustain a growth rate of eight to nine per cent, which could be even higher if it were to introduce some key reforms. In terms of growth in the factors of production, gross investment in India has risen from 25 per cent of the GDP in 2002–03 to 36 per cent in 2006–07, and India's population is predicted to become on average younger, implying faster growth in the workforce. The higher proportion of the workforce in the population also promises to raise savings and investment further. As for productivity growth, the competitive pressures on entrepreneurs brought about by the external and internal opening up are here to stay. The changes in the initial conditions brought about by the structural changes in the post-reform era offer yet another reason to take an optimistic view.

Reforms and Growth

That the command and control policies served India rather poorly is not very much in dispute. It is generally agreed that the country's economic performance in Phase II was quite poor. Those who spent time in India during this period would testify to very little change in the country. Poor performance was reflected most visibly in scarcities and poor product quality. People who wanted a scooter, automobile or telephone had to wait for a year or longer. Phone service was so poor that half the time people did not get a dial tone, and when they did, they were frequently connected to a wrong number. Bicycles in the 1970s were hardly any different from those in the 1950s. The same held true of automobiles.

There is less agreement on the 1980s and beyond, however. DeLong (2003) initially raised the question by arguing that growth in India had accelerated in the 1980s prior to the reforms of the 1990s. Building on this argument, Rodrik (2003) raised the stakes, asserting that the 'change in official attitudes in the 1980s' may have had a bigger impact than any specific policy reforms. Panagariya (2004) questioned this assertion, arguing that:

- piecemeal liberalizing reforms had already begun in the late 1970s and continued through the 1980s
- but for the super-high growth during 1988–91, the last three years of the decade, the growth rate in the 1980s was significantly lower than in the 1990s
- this super-high growth rate was partially fuelled by fiscal expansion which could not

be sustained, as evidenced by the 1991 financial crisis
- regardless of the trigger, the higher growth rate could not have been sustained without liberalizing reforms.

Panagariya (2008) also argues that the recent acceleration of the growth rate to nearly nine per cent further strengthened the argument that reforms were critical to accelerating and sustaining high growth rates.

Three examples may help to buttress the argument that without the liberalizing reforms of the 1990s and beyond, India could not have achieved its transformation. First, India's exports of goods and services as a proportion of GDP rose from 7.3 per cent in 1990–1 to 13.6 per cent in 2002–03 and 21 per cent in 2007–08. If India had kept blanket licensing on virtually all imports, and the high tariffs which averaged 113 per cent and peaked at 355 percent in 1990–1, it is inconceivable that this tripling of the ratio could have happened. Second, in 1990–1, foreign investment inflow into India was a paltry US$6 million. It rose to $6 billion in 2002–03 and $61.8 billion in 2007–08. Without liberalization, this change would have been impossible. Finally, tele-density (phones per 100 population) rose from less than three in 1998 to 31 in October 2008. The total number of phones was less than 6 million in 1990–1. It rose to 76.3 million in 2002–03 and to 364 million at the end of October 2008. Even rural India could boast of 109 million phones by October 2008. Without the telecoms sector reforms of the 1990s and 2000s, this expansion would also have been impossible. These examples lead to the conclusion that had India heeded those who recommended a move away from a command and control regime, including an end to import licensing (as in the pioneering work of Bhagwati and Desai 1970), it would have reached a higher growth path much sooner.

Poverty and Inequality

There has been a vibrant debate on whether the reforms and accompanying growth have led to poverty reduction in India. In part, this debate was fuelled by a change in sample design of the large sample survey in 1999–2000, which made it noncomparable to the preceding large survey done in 1993–4. This debate was resolved by another large survey in 2004–05. There is now general agreement that while the precise decline in poverty depends on where one draws the poverty line and which price index is used to convert the poverty line from one year to another, significant reduction in the poverty ratio, measured as the percentage of poor people to the total population, did take place between 1993–4 and 2004–05. Significant reduction in the poverty ratio also took place between 1983 and 1993–4. According to the official calculations by the Planning Commission, the poverty ratio fell from 44.5 per cent in 1983 to 36 per cent in 1993–4 and 27.5 per cent in 2004–05. In contrast, there was no change in the trend poverty ratio between 1951–2 and 1973–4.

Reform critics also argue that post-reform growth has led to massive inequalities in the country. If we use the conventional Gini coefficient as the measure, there is no perceptible increase in its value between 1983 and 2004–05. But regional inequality, as measured by state-level per-capita incomes, and urban–rural inequality, has gone up. This should not be surprising. Rapid growth often creates urban agglomerations which concentrate in a few regions, and therefore leads to both regional and urban–rural inequality. A little-appreciated fact is that even South Korea, which is often cited as an example of rapid growth with equity, actually experienced both regional and urban–rural inequality during its rapid-growth phase (Ho 1979).

A danger of excessive focus on inequality is that it can lead to policies that undermine wealth creation, growth and ultimately poverty alleviation. India's own experience in the second half of the 1960s and 1970s demonstrates the dangers of this approach. Concerns with 'concentration of wealth' largely motivated Mrs Gandhi to impose severe controls on investments by large firms and big business houses, marginal income-tax rates that exceeded 95 per cent, and the reservation of labour-intensive products for SSEs. These

policies scuttled growth and any hope of helping the poor.

A more effective way to combat regional and urban–rural poverty is to concentrate government efforts on anti-poverty programmes. In so far as the poor are concentrated in low per-capita-income states such as Bihar, Orissa and Uttar Pradesh, and in rural areas, efforts to fight poverty will automatically help reduce regional and urban–rural inequalities.

India's Challenge

A key feature of India's growth, different from almost all other countries at a similar stage of development, is the disappointing performance of manufacturing. Whereas a rapidly rising share of industry in the GDP accompanied growth in South Korea and Taiwan in the 1960s and 1970s and in China more recently, the same has not happened in India. The share of industry (including manufacturing; mining and quarrying; and electricity, gas and water) in India's GDP has remained remarkably steady at around 21 per cent since 1990–1. The decline in the share of agriculture and allied activities during these years, from 29.3 per cent in 1990–91 to 17.8 per cent in 2007–08, has been entirely absorbed by services, whose share has grown from 49.2 per cent to 61.4 per cent over the same period.

Although formal sectors such as information technology, telecommunications and finance have shown rapid growth, services largely consist of informal sector services. Within industry, labour-intensive products such as apparel, footwear, toys and other light manufactures that generate well-paid jobs have done poorly. This has meant that the creation of well-paid jobs in the economy has lagged despite rapid growth.

To put it differently, approximately three-fifths of the workforce currently derive their income from agriculture and allied activities, while these sectors generate less than one-fifth of the total income. Given that agricultural growth rarely exceeds four per cent, this means that an extremely large part of the population is not sharing in India's rapid growth. In part, this calls for reforms in agriculture to accelerate growth in that sector. But more importantly, India needs to create well-paid jobs in industry and services far more rapidly in order to pull a large chunk of the labour force out of agriculture. In a modern economy, agriculture can be the primary source of income for only a small proportion of the population. If a substantial proportion of the agricultural workforce migrated to different sectors, this would will also reduce the pressure on land and raise agricultural productivity.

A key reform necessary to accelerate job creation in the formal industrial sector relates to the labour market. Firms must have the right to terminate workers upon payment of appropriate severance pay. The current regulations have worked as a serious barrier to the entry of large-scale firms in the labour-intensive sectors in India. Apparel factories in India have tended to be much smaller than even in Bangladesh and Sri Lanka. With labour accounting for the bulk of the costs in the labour-intensive sectors, firms are reluctant to enter these sectors on a large scale in the absence of the right to terminate workers.

What Can We Learn from the Indian Experience?

India's rich experience offers several lessons. First and foremost, it shows that democracy is not a barrier to rapid growth. Until recently, analysts argued that democracy might be consistent with growth rates of from four to six per cent, but not much higher rates. Even Chile, which is a democracy, has not been able to break the six per cent barrier on a sustained basis. India has now grown at almost nine per cent for five years, and despite the current hiccups, it promises to maintain that rate in the second decade of the 21st century.

Second, reforming a highly distorted economy is a long-drawn-out process. Some analysts have recently argued that countries should look for one or two policies that most constrain growth, and concentrate on changing them. India's experience demonstrates otherwise: the reform process extends to decades. Relaxing constraints on growth in one area only expose the constraints in

other areas. Success on a sustained basis requires sustained action over several decades.

Third, low or declining barriers to trade are extremely critical to rapid growth. Side-by-side, it is necessary to give entrepreneurs space in which they may freely operate. Reforms necessary to give entrepreneurs the necessary space may vary from country to country, but at the end of the day, each country must find ways to free up entrepreneurs to seek profits without undue restraints.

Finally, the country must own its own policies. Forced policy reform by the World Bank and the International Monetary Fund (IMF) in the 1980s in many countries was destined to fail. When governments themselves do not own a reform, they will sabotage it at the implementation level or reverse it once the loan has been disbursed. India's reform was fully owned by its successive governments. At worst, we could argue that the first set of reforms in 1991–2 were carried out under the terms and conditions of the IMF and the World Bank loans. But even this is contestable, as is pointed out in chapter 5 of Panagariya (2008). Everything else that followed was initiated and executed by India. After India's 1991–92 programme, the IMF became irrelevant to the country. Likewise, following the first structural adjustment loan by the World Bank, India was firmly in the driving seat. The World Bank remained engaged only because it wanted to loan money to India, and did so by selling the loans internally to its board based on the policy actions India had been taking. Without ownership, reforms would not have been sustained. Nor would they have been credible to the entrepreneurs.

See Also

► India, Economics in

Bibliography

Bhagwati, J., and P. Desai. 1970. *India: Planning for industrialization*. London: Oxford University Press.

DeLong, J.B. 2003. India since independence: An analytic growth narrative. In *In search of prosperity: Analytic narratives of economic growth*, ed. D. Rodrik. Princeton: Princeton University Press.

Ho, S.P.S. 1979. Rural–urban imbalances in South Korea in the 1970s. *Asian Survey* 19(7): 645–659.

Maddison, A. 1971. Ch. 5: Reasons for the acceleration of economic growth since independence. In *Class structure and economic growth: India and Pakistan since the Moghuls*. London: George Allen and Unwin.

Nehru, J. 1946. *The discovery of India*. New York: John Day.

Panagariya, A. 2004. Growth and reforms during 1980s and 1990s. *Economic and Political Weekly* 2581–2594.

Panagariya, A. 2008. *India: The emerging giant*. New York: Oxford University Press.

Rodrik, D. 2003. Institutions, integration, and geography: In search of the deep determinants of economic growth. In *In search of prosperity: Analytic narratives of economic growth*, ed. D. Rodrik. Princeton: Princeton University Press.

I

Indian Economy: Yesterday, Today and Tomorrow

Arvind Panagariya

Abstract

This article offers an analytic overview of India's achievements to date, what its future prospects are, what its rise means to the global economy in the next fifteen years and what challenges India faces in terms of future reforms. The article begins by presenting a summary of the country's growth experience during the last sixty years and relating it to the policies and political economy factors behind the adoption of those policies. It then discusses medium-term prospects of the country. Based on a set of key factors relevant to growth, it argues that India is likely to become the third largest economy in the world and an even bigger contributor to the global workforce than it is today. The article then turns to the study of the impact the growth has had on poverty alleviation during India's sixty-year history. The remainder of the article outlines the key challenge India faces today and the reforms it needs to undertake to sustain and accelerate both growth and poverty alleviation. The article argues that India needs to walk on

two legs – manufacturing and services – and requires reforms that would help strengthen both.

Keywords
Challenge; Global economy; Growth; India; Poverty; Redistribution; Reforms

JEL Classifications
O11; O14; O19

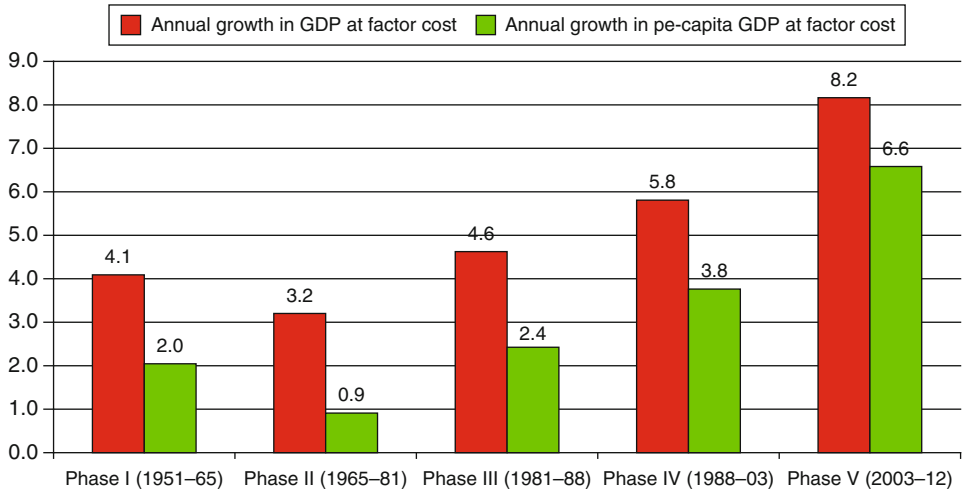
The Indian economy grew an impressive 8.5% per year between the financial years 2003–04 (1 April 2003 to 31 March 2004) and 2010–11. This period included the global-financial crisis year of 2008–09. Unlike most other economies around the world, the Indian economy was barely dented by the crisis. It experienced a small and temporary decline in the growth rate to 6.8% during 2008–09 and bounced back to the 8% plus rate in the following two financial years. This makes the compelling point that the economy has shifted onto a tiger-like growth trajectory for some years to come.

Yet, with surprising speed, pessimism bordering on the gloom of the 1970s and 1980s has returned to the Indian economic scene. The growth rate during the last three-quarters of the financial year 2011–12 fell progressively to 6.7%, 6.1% and 5.3% and has recovered only marginally to 5.5% in the first quarter of 2012–13. The growth rate in 2011–12 as a whole was the lowest, at 6.5%, since the economy began growing at an 8% plus rate in 2003–04. The decline has been accompanied by high inflation, large and rising fiscal deficit and a depreciated rupee against the dollar. The result has been a certain degree of panic among commentators on the Indian economy, with some predicting that the 'I' is about to drop out of the acronym BRIC, others suggesting that the only vowel in the acronym now stands for Indonesia and still others pronouncing an end to India's growth story (see, for example, the article by Mukherji and Ogawa (2012) of Standard and Poor entitled 'Will India be the first fallen BRIC angel?'. The article goes so far as to raise the possibility that India may revert to the pre-reform growth rate of 4–5%. To quote from its concluding section, 'Some observers in India possibly assume that the economy could sustain 6%–7% GDP growth in the coming years without active reforms or more effective economic management. However, we should not exclude the possibility of a more significant drop in trend GDP growth (perhaps to 4%–5%) if weak economic management coincides with a bad external shock or with bad luck, such as a poor monsoon'.).

Against this background, the present article offers an analytic overview of what India has achieved to date, what its future prospects are, what its rise means to the global economy in the next fifteen years and what the challenges are that India faces in terms of future reforms. The first section begins with a summary of the growth experienced during the last sixty years, the role of economic policies in determining this growth trajectory and the political economy factors that led to the adoption of the policies. The second section discusses the medium-term prospects of the country. Based on a set of key factors relevant to growth, it is argued that India is likely to become the third largest economy in the world and an even bigger contributor to the global workforce in the next fifteen years. The third section turns to the impact that growth, or lack thereof, has had on poverty alleviation during the sixty-year history. The fourth section outlines the key challenge that India faces today. The fifth section turns to the reforms that India needs to sustain and accelerate both growth and poverty alleviation. It is argued that India needs to walk on the two legs of manufacturing and services, and requires reforms that would help strengthen both.

Growth: An Overview

It is useful to divide approximately sixty years worth of modern economic history of India into five separate phases, as done in Fig. 1. (The rationale for the phases chosen in Fig. 1 is provided in Panagariya (2008, Chapter 1), which also offers a detailed account of the twists and turns in India's economic policies.) After independence, India



Indian Economy: Yesterday, Today and Tomorrow, Fig. 1 Annual growth rates during five phases, 1951–52 to 2011–12 (Source: based on the author's calculations using the data in the *Handbook of Statistics on Indian Economy, 2012* by the Reserve Bank of India at http://www.rbi.org.in/)

launched its development programme, with the First Five-Year Plan beginning in 1951–52. (Although planning as the principal tool of development has now been largely abandoned, India continues to use the five-year plans as a major medium-term policy statement. Because of breaks between the end of one five-year plan and the launch of the next several on several occasions, India will be launching its Twelfth Five-Year Plan in 2012–13.) In conformity with the prevailing conventional wisdom among all economists, western and Indian, the state was given a key role in the development process. Two objectives guided the policy: self-reliance, interpreted as the absence of dependence on imports to satisfy domestic demand and on exports for the sale of goods produced at home, and therefore a coincidence of production and consumption baskets; and a steady increase in the share of the public sector in investment and output. With the national interest rather than that of a colonial power placed at the centre of the development effort, institutions of a vibrant and durable democracy were put in place: a British-style parliament to legislate, a fiercely independent judiciary, a free press and a substantial bureaucracy headed by the country's brightest men and women. The result was satisfactory though not spectacular economic growth: during the first 14 years of planned development, ending with the year 1964–65, GDP at factor prices grew at the annual rate of 4.1%. Allowing for population growth of 2.1% per year, per capita GDP growth during the period was 2% per annum.

The first 14 years, or Phase 1, largely coincided with the rule of Prime Minister Jawaharlal Nehru. Two key instruments were deployed during this period to catalyse growth while promoting self-reliance by progressively realigning the production basket to the one domestically consumed and expanding the public sector. First, the public sector entered production activity in heavy industry sectors, such as steel and machinery, on the premise that the private sector lacked the resources required for investment in them. Second, larger investment projects in the private sector were subjected to licensing to ensure that private investments were channelled into highpriority sectors rather than being guided purely by profitability. The system worked relatively smoothly through the 1950s. During that decade, barriers to trade were also low, with many consumer goods imports permitted.

A balance of payments crisis in 1957–58 led the government to adopt foreign exchange budgeting under which the finance ministry would predict the expected foreign exchange revenues in the following six-month period and

allocate it across its competing uses. This naturally created an extra layer of bureaucracy in the investment licensing system, since a licence could not be issued unless foreign exchange for the machinery and raw materials necessary for the project was available. This factor, complemented by a rising volume of private investment, which translated into progressively larger number of applications for licences, began to create serious bottlenecks in the administration of the licensing system as the Nehru era ended. Beginning in 1964, several committees were set up to recommend changes that would help streamline the licensing procedures, but the efforts were largely unsuccessful.

Nehru died in 1964 and was succeeded by Prime Minister Lal Bahadur Shastri, who was more sympathetic to agriculture and unenthusiastic about heavy industry. His administration oversaw the launch of the Green Revolution and the creation of much of the infrastructure of public distribution system of food grains, including the Food Corporation of India (FCI) and Agricultural Prices Commission which still exist today. Unfortunately, however, Shastri died unexpectedly in early January 1966 and was succeeded by Prime Minister Indira Gandhi, daughter of Nehru.

Political compulsions led Gandhi to turn to a far more extreme form of socialism than under Nehru. Investment by a large firm or business house, formally defined as a firm or interconnected groups of firms with 350 million rupees or more in assets, was confined to 19 highly capital-intensive core industrial sectors. Alongside, most of the labour-intensive sectors, such as apparel, footwear and light consumer goods of all kinds, were reserved for exclusive manufacture by small-scale enterprises. Imports were subjected to such tight controls that they fell to just 4.1% of GDP in 1969–70. Foreign investment rules, which had been relatively liberal under Nehru, were tightened dramatically. With some exceptions, foreign companies were told either to register as Indian companies or to leave the country. Two major American companies, IBM and Coca-Cola, left India in the second half of the 1970s. The largest 14 domestic banks, insurance companies, coal mines and oil companies were nationalised. A ceiling was placed on urban land holdings. Land held above the specified ceiling was to be put for sale, with the government having the right to buy it at a throwaway price. The result was the disappearance of much urban land from the market. Finally, labour laws were changed to further favour workers, with manufacturing firms employing 100 or more workers denied the right to lay off workers under any circumstances.

These draconian measures had a chilling impact on the economy. While industrial countries boomed during the 1960s and early 1970s (until the oil-price crisis put the brakes on that growth), and countries such as South Korea and Taiwan that had chosen an outward-oriented path to development went on to achieve growth rates ranging from 8% to 10%, the growth rate in India dipped. The decade from 1965–66 to 1974–75 produced GDP growth of just 2.6% per year, with per capita GDP rising just 0.3% annually. This was a lost decade for the country.

By the second half of the 1970s, at least some in the government began to recognise that the controls had gone too far for the good of the economy. Although this was not publicly acknowledged and no policy change was actually announced, some piecemeal liberalisation involving the expansion of production capacity under the existing licenses and freer imports of machinery and raw materials was introduced. This process accelerated in the 1980s, especially in the second half of the decade under Prime Minister Rajiv Gandhi, who succeeded his mother following her assassination in 1984. Growth recovered to 4.2% during 1975–81 (1975–76 to 1980–81) and to 4.6% during 1981–88.

Although the small acceleration in growth during the 1980s was partially stimulated by the piecemeal reforms, it was also fuelled by an expansionary fiscal policy that relied on substantial overseas borrowing. As the 1980s closed, this borrowing had led to an accumulation of substantial external debt. Moreover, despite an acceleration of export earnings in the second half of the 1980s, due to significant depreciation of the rupee and the introduction of some export incentives, their level remained low. As a result, debt

servicing came to absorb nearly 30% of the meagre export earnings by the end of the 1980s. In turn, foreign exchange available for imports progressively dwindled and a balance of payments crisis followed in 1991.

This crisis coincided with an election campaign during which Rajiv Gandhi was assassinated, paving the way for Prime Minister Narasimha Rao to take the helm. Contrary to his reputation, Rao proved to be a decisive prime minister who used the occasion of the crisis to introduce major reforms. He abolished licensing on investment and imports of capital goods and raw materials (though not on consumer goods). He also opened the economy to foreign investment. In the subsequent years, he extended the reforms to telecommunications, civil aviation and the financial sector, while continuing to liberalise trade through tariff reductions.

Rao lost the election in 1996 and was followed by three short-lived coalition governments. Eventually, in 1998, Prime Minister Atal Bihari Vajpayee came to head the Bharatiya Janata Party (BJP) led National Democratic Alliance (NDA) government. Though Vajpayee lost a crucial vote in his first year, leading to the fall of his government, the electorate returned him with a stronger mandate in 1999. He ruled until May 2004, when the NDA lost a crucial election.

The reform process not only continued but also accelerated considerably under Vajpayee. Import licensing on consumer goods imports was ended and tariffs were systematically brought down, with the highest tariff on industrial goods dropping to just 10% (with some exceptions) in the last budget presented under this government. A major reform of the telecommunications sector paved the way for fierce competition among private and public providers. The result was an explosion in the growth of mobile phones in India. Major initiatives were also undertaken in the area of infrastructure, including the building of highways and rural roads and modernisation of ports. Other reforms included the liberalisation of interest rates, freer entry to domestic private and foreign banks, freeing up of markets in agricultural produce, entry of the private sector to insurance, steady trimming of the list of sectors subject to small-scale industries reservation, privatisation of several public-sector enterprises and the repeal of the central urban land ceilings act, which paved the way for state governments to drop the ceiling on urban land holdings. These changes went a long way towards intensifying competition in various markets.

The reforms under the Rao and Vajpayee governments went a long way towards accelerating growth. Although growth had crossed the 7% mark during the three years preceding the 1991 crisis, it could not be sustained due to its partial origins in the expansionary fiscal policies. But the higher growth during the 1990s followed fiscal consolidation and pro-market reforms. As a result, it was not only sustained but accelerated. India grew 5.8% per annum during 1988–89 to 2002–03 and then, in 2003–04, shifted to the higher growth path of 8–9%. Although the growth rate fell to 6.8% in 2008–09 following the global financial crisis, it quickly returned to the 8% plus range in 2009–10 and 2010–11.

The Congress returned to power in May 2004, heading a coalition that came to be known as the United Progressive Alliance (UPA) and consisted of approximately a dozen large and small parties. Dr Manmohan Singh, an economist who had served as the Finance Minister in the Rao government and had guided the reforms in the first half of the 1990s, was appointed Prime Minister. Unfortunately, however, this government interpreted the defeat of the NDA as a vote against reforms. It proclaimed its intention to promote reforms with a 'human face'. In effect, this rhetoric translated into an end to pro-growth reforms and to attention being focused nearly exclusively on redistributive programs. Even progress on building the country's infrastructure slowed down. Perhaps the most visible policy initiative of the government was the introduction of a large-scale National Rural Employment Guarantee Scheme under which one member of each rural household is guaranteed employment for 100 days at a wage significantly above the equilibrium rate.

While the UPA government did not follow up on the Rao–Vajpayee reforms, it generally did not do anything significant to impede their effects from being realised. As a result, the growth

acceleration that had taken place in 2003–04 was sustained during its rule. This steady growth of 8–9%, in turn, helped the government return to power in 2009. (Gupta and Panagariya (2012) provide an empirical analysis of the 2009 election and show that growth is the key to explaining the outcome in this election.) But the policy environment began to deteriorate during the second term of the UPA.

Two factors, in particular, hurt the economy. First, in an overreaction to high inflation, the Reserve Bank of India curbed the growth of the money supply through 13 consecutive increases in the interest rate. The resulting increase in the cost of funds had an adverse impact on private investment. In parallel, high fiscal deficits had the obvious effect of crowding out some private investment.

Secondly, and more importantly, almost from the beginning of the second term of the UPA a paralysis gripped the administrative and policy-making processes of the government. The paralysis began with a hyperactive environment minister blocking clearances to hundreds of projects around the country. Later, revelations of a large number of corruption scandals, followed by the imprisonment of two ministers, one Member of Parliament and several civil servants, led to a chill in the entire decision-making machinery. Civil servants would no longer take action on the basis of verbal orders by their ministers, while the latter came to fear issuing even legitimate orders in writing lest they were accused of doing so in return for a bribe.

This paralysis in decision-making in the administrative machinery has been accompanied by paralysis in policy making. The Prime Minister and his Congress party have been utterly unsuccessful in negotiating policy changes with their coalition partners. One or the other coalition member has gone on to block every important policy initiative of the government. The result has been legislative paralysis as well.

These two factors have been largely behind the recent growth slowdown. While the Reserve Bank of India is beginning to reverse its tight monetary policy, its ability to continue doing so is constrained by fiscal deficits. Accommodation of large fiscal deficits by the Bank inevitably risks fuelling inflation. At the same time, the paralysis in administrative and policy-making processes has not gone away either. Without corrective action by the government on both fronts, a perfectly plausible growth story runs the risk of being stopped dead in its tracks.

Medium-Term Growth Prospects

Setting aside these two considerations, the prospects for growth in India over the next fifteen years are excellent. To fully appreciate this fact and its implications for India's position in the global economy, consider first the growth that India has achieved during the last nine years in US dollar terms. The simple average of annual growth rates in current dollars during the nine years beginning in 2003–04 and ending in 2011–12 has been 15.8%. Even allowing for 3% per year inflation in the USA, this figure implies a growth rate of 12.8% in real dollars.

Making next the conservative assumption that the GDP in real dollars will continue to grow 10% per year in constant dollars, it will expand from a GDP of 1.8 trillion dollars in 2011–12 to 7.5 trillion in 2025–26 in 2011–12 dollars. India would then become the third largest economy in the world after the USA and China. Moreover, even applying the current population growth rate of 1.8% per annum, which is bound to decline, this GDP will imply a per capita GDP of $4,800 at 2011–12 prices. That would spell the end of poverty as currently defined with near certainty.

Demographically, the number of workers aged 20 to 49 years is predicted to decline by 37 million in developed countries and 63 million in China between 2010 and 2025. In India, this number is predicted to increase by a gigantic 131 million. With an increased international mobility of workers, these numbers are likely to translate into young Indian workers becoming far more ubiquitous around the globe than today. Rising incomes within India, which would make it possible for parents to send their children abroad for education, will only facilitate this process of emigration.

The critical question to which we must return is why the prospects of 10% growth in real dollars are good for the next 15 years. At least four factors allow us to make a compelling case. First, investment rates in India have hovered around 35% during the last several years. This investment is largely financed by domestic savings, which means that the savings rate has also been 30% or more during these years. Going by the historical experience of countries such as South Korea, Taiwan, Singapore and China, it is quite unlikely that these savings and investment rates will collapse in the near future. Given that India is predicted to become progressively younger, labour shortages will not act as a brake on growth either.

Second, the reforms introduced during the 1990s and early 2000s have remained intact. This means that India remains a highly open economy in at least industrial goods and services, although it is still highly protected in agriculture. Domestic entry is also relatively free. Therefore, entrepreneurs must compete intensely with one another. Large inefficiencies remain in check.

Third, complementing this competition effect is the gap between productivity in India and that in the 'best-practice' countries. This large gap offers India significant scope for technological catch-up. Therefore, in addition to growth through increased factor supply (capital and labour), India has the possibility of adding to its growth rate through productivity gains. Because entrepreneurs are subject to intense competitive forces, it is likely that such productivity gains through technological advance will be realised.

Finally, rapid growth also requires entrepreneurs willing to take risks. Luckily here as well India has had a longstanding tradition of entrepreneurship. The fact that even during the age of licensing, with both hands virtually tied, Indian entrepreneurs could produce 3–4% growth almost steadily is evidence of their skills and talent. Surely, in the current reformed environment and with further reforms likely in the forthcoming years, the chances of their performance improving yet more are very good.

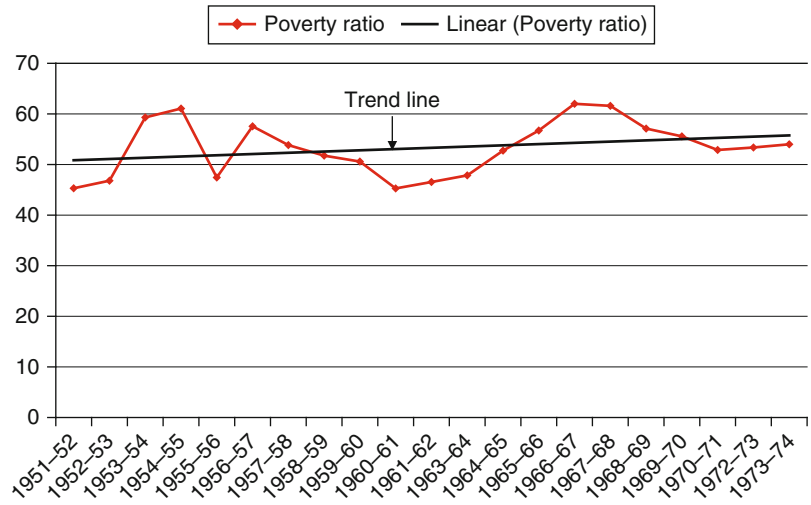
Poverty and Inequality

An issue of great importance is the impact the reforms and growth have had on poverty. Evidence is now compelling that growth has not only been accompanied by a decline in poverty, but that the acceleration of this growth has been accompanied by an accelerated decline in poverty.

In 1950–51, India started at an extremely low per capita income and a large percentage of the population was in poverty at the beginning. But the country also grew slowly during the first 25 years. That in turn meant that even in the mid-1970s, per capita incomes remained quite low and no progress could be made in poverty alleviation. With such low per capita income, even the scope for poverty alleviation through redistribution was extremely limited. It was only when growth accelerated and per capita income began rising more rapidly that an impact on poverty was discernible. Not only did growth begin to 'pull up' people into gainful employment, it also produced larger tax revenues that could be used to finance enhanced redistribution programs, such as the National Rural Employment Guarantee Scheme.

Before presenting the evolution of poverty, it should be noted that there has been some controversy in India about the level at which official Indian poverty lines are set. The original official rural and urban poverty lines, known as the Lakdawala poverty lines, had been set at the recommendations of a 1993 expert group headed by the then leading poverty expert Professor D. T. Lakdawala. Although the Planning Commission recently revised the rural poverty line upwards at the recommendation of the Tendulkar Committee report, thereby aligning it to the urban poverty line in real terms, many have argued that the poverty line still remains low. My own view is that in a country with widespread poverty, an important role of the poverty line is to allow tracking of the fortunes of those living in destitution. Therefore, setting this line near the subsistence level has some merit.

Whatever one's view with respect to the level at which poverty line is set, if we are interested in comparable estimates for the entire sixty-year period under consideration we are confined to



Indian Economy: Yesterday, Today and Tomorrow, Fig. 2 Poverty ratio, 1951–52 to 1973–74 (Source: author's construction using data from Datt (1998, Table 1))

the Lakdawala poverty lines. Even at the Tendulkar lines, estimates go only as far back as 1993–94. In principle, it is possible to extend the estimates at the Tendulkar lines back in time or to calculate them at alternative lines, but they are not readily available.

Accordingly, the estimates presented here are based on the Lakdawala lines. Figure 2 shows the evolution of the poverty ratio – the proportion of the population below the poverty line – from 1951–52 to 1973–74. As is readily seen from the trend line, there was no long-term reduction in poverty during this period. Indeed, the trend line shows a slight upward movement over time. Neither was growth robust enough to pull people out of poverty nor did the meagre tax revenues generated by low levels of income provide enough resources for significant redistribution. But this changed as growth picked up and income began to rise.

Table 1 reports the poverty ratio in rural and urban India and the two regions taken together approximately every five years beginning in 1973–74. (There is one exception since the estimates jump to 2004–05 after 1993–94. Although a thick expenditure survey was conducted during this year, due to a change in the sample design estimates based on it are not strictly comparable to those in the other years. Therefore the estimates associated with this year have been suppressed.) These estimates are based on the so-called thick surveys that typically collect expenditure data on over a hundred thousand households nationwide. As is readily seen, the acceleration in growth beginning in 2003–04 also translates to accelerated poverty reduction. For example, reduction in total poverty was 0.77 percentage points per year from 1993–94 to 2004–05 but 1.44 percentage points per year from 2004–05 to 2009–10. Mukim and Panagariya (2012) provide a comprehensive analysis of the evolution of poverty in India. They show that poverty has fallen steadily since 1983 for all major social and religious groups and states. There is simply no truth in the common assertions that growth has impoverished

Indian Economy: Yesterday, Today and Tomorrow, Table 1 Poverty ratio, 1973–74 to 2009–10 (Source: Planning Commission for estimates until 2004–05 and Mukim and Panagariya (2012) for 2009–10)

Year	Rural (%)	Urban (%)	Total (%)
1973–74	56.4	49	54.9
1977–78	53.1	45.2	51.3
1983	45.6	40.8	44.5
1987–88	39.1	38.2	38.9
1993–94	37.3	32.4	36
2004–05	28.3	25.7	27.5
2009–10	20.2	20.7	20.3

the socially disadvantaged or that it has failed to benefit specific religious minorities.

Recent research also shows that contrary to common assertions that growth has increased inequality in India, no unique relationship is observed between these two variables. Krishna and Sethupathy (2012) measured inequality using the Theil index. This index allows them to distinguish between within group and between group inequalities. Using the expenditure data from surveys conducted in 1987–88, 1993–94, 1999–2000 and 2004–05, they show that the overall inequality shows only modest variation over the period. It rises slightly between 1987–88 and 1993–94 and again between 1993–94 and 1999–2000, but falls by 2004–05 to roughly the 1987–88 level.

Hnatkovska et al. (2012a) also show that the gaps in wages and education levels between scheduled castes and tribes on the one hand, and non-scheduled-caste groups on the other, have steadily declined between 1983 and 2004–05 (scheduled castes and scheduled tribes refer to historically socially disadvantaged groups in India). The gaps exhibit a decline when measured using mean and median wages and education levels over time as well as when evaluated in terms of intergenerational mobility rates. Indeed, scheduled caste and scheduled tribe children have changed their status relative to their parents in terms of wages and education even faster than non-scheduled caste children between 1983 and 2004–05 (Hnatkovska et al. 2012b).

Regional inequality, rural–urban inequality and inequality between the richest and the poorest (however defined) have certainly risen. But a moment's reflection will show that these forms of inequality are nearly impossible to escape in fast-growing economies and have been a part of all growth miracles, such as South Korea and Taiwan in the 1960s and 1970s and China more recently. Growth involves wealth generation, and those creating wealth are bound to end up with at least a small part of it for themselves, while the remainder is distributed over the rest of the population. And when the wealth generated runs into tens of billions of dollars, even a small fraction of it is a lot of wealth for a single individual. This fact alone suffices to raise the inequality between the richest and the poorest. Likewise, fast growth concentrates in a small number of agglomerations that are located in urban areas; even if they begin in rural locations, the growth turns them into urban areas over time. This pattern necessarily leads to regional and rural–urban inequality.

The Challenge Facing India

A key difference between India and other fast-growing economies, such as South Korea and Taiwan in the 1960s and 1970s and China more recently is that poverty reduction per percentage-point growth in India has been significantly smaller. Whereas two decades of rapid growth in these other countries was sufficient to wipe out abject poverty, this has not been the case in India. By all measures, at least a fifth of the Indian population still lives in what is sometimes called extreme poverty.

This slow progress in combatting poverty has in turn been due to a development pattern that is so far unique to India. In almost all cases of rapid growth in labour-abundant developing countries, manufacturing in general and labour-intensive manufacturing in particular have led the process. In turn, this process has allowed the countries to shift the workforce rapidly out of agriculture and into well-paid jobs in manufacturing activities while also fuelling urbanisation. In addition to providing gainful employment to those migrating from the countryside, this process has also helped raise output per worker in agriculture by reducing the land-to-workers ratio.

Unfortunately, capital-intensive and skilled-labour-intensive manufacturing and services sectors have led the growth process in India. The successful sectors in India are telecommunications, information technology, automobiles, motorbikes, petroleum refining, finance and pharmaceuticals. Labour-intensive sectors such as apparel, footwear and light consumer goods manufacture have not flourished. For example, apparel exports from India are less than those from Bangladesh and one-tenth of those from China. This pattern has meant that while the

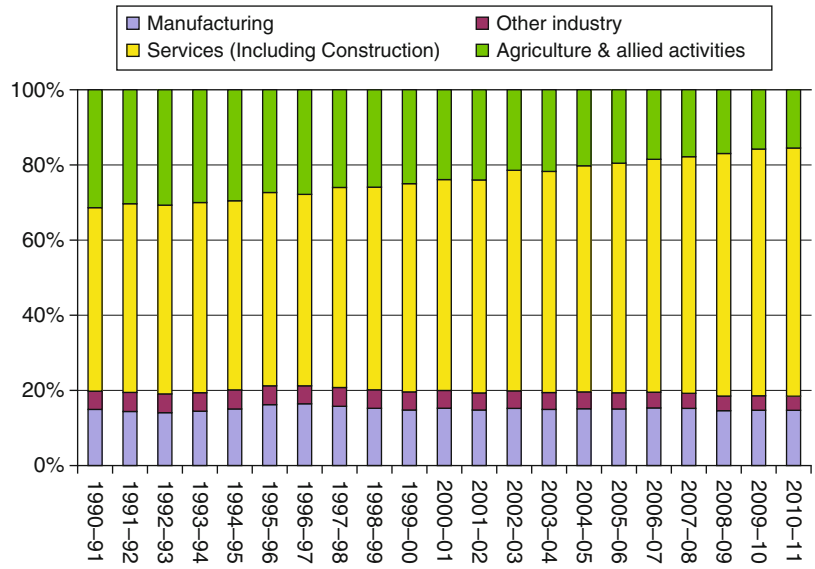
share of agriculture in the GDP has significantly declined, just as it did in South Korea and Taiwan in the 1960s and 1970s and China more recently, the employment share of agriculture has fallen more gradually (see Table 2). Indeed, until recently, the absolute number of workers in agriculture has continued to grow. The net effect of this pattern has been slow growth in gainful employment in industry and services, slow growth in output per worker in agriculture and slow pace of urbanisation. All of these factors have had a dampening effect on the pace of poverty reduction.

For a labour-abundant country like India, with more than 500 million workers, the greatest potential comparative advantage lies in labour-intensive manufacturing. With specialisation in these products, it could exploit the vast world markets. The same opportunities do not exist in the capital- and skilled-labour-intensive products. As a result, specialisation in the latter, being at least partially limited by the size of the domestic market, has meant that manufacturing as a whole has grown far slower than in other successful labour-abundant economies. This fact is clearly illustrated by Fig. 3, which shows the evolution of the shares of the major sectors of the economy during the post-reform era. While agriculture has rapidly lost share in the GDP, the shares of manufacturing and other industry sectors have remained essentially unchanged. Services have in turn taken up the slack, expanding their share of GDP.

From a good jobs perspective, this pattern would not be all bad if services had generated a

Indian Economy: Yesterday, Today and Tomorrow, Table 2 Pattern of development in the post-reform era (Source: author's calculations using the data from the *Handbook of Statistics on Indian Economy, 2012* by the Reserve Bank of India at http://www.rbi.org.in/, reports on employment–unemployment survey by the National Sample Survey Organization, various years, and the Census of India, various years)

Item	1993–94	2004–05	2009–10
GDP share of agriculture and allied activities	30	20.2	15.8
Share of agriculture and allied activities in rural workforce	78.4	72.7	68
Urbanisation (nearest census: 1991, 2001 and 2011)	25.7	27.8	31.2



Indian Economy: Yesterday, Today and Tomorrow, Fig. 3 Composition of GDP, 1990–91 to 2008–09

large number of formal sector jobs. But this has simply not been the case. The work by Dehejia and Panagariya (2016) shows that almost three-fifths of services sector workers are employed in own-account enterprises that employ no hired workers on a regular basis. The small proportion of workers that has moved out of agriculture has largely ended up in low-paid informal sector jobs. The critical questions are why has the pattern of development in India been so different and what can be done to stimulate faster growth in well-paid jobs?

Walking on Two Legs: What India Must Do

Because half of the workforce in India still remains in agriculture and lives on a meagre income, it is tempting to conclude that improving incomes in that sector can bring the fastest relief to the poor. Yet this is somewhat misleading, since growth in agriculture has rarely exceeded 4% per year on a sustained basis in India. Therefore, while there is merit in doing what can be done to improve agricultural incomes, there is no alternative to creating well-paid jobs in industry and services, paving the way for a rapid movement of workers out of agriculture into these sectors. Indeed, such movement will contribute the most to increased incomes in agriculture by rapidly bringing down the worker-to-land ratio.

Even though past experiences have seen manufacturing largely leading the growth process, given that India has already achieved some success in the services sector it makes more sense to now walk on two legs: manufacturing and services. Future reforms must focus both on stimulating labour-intensive manufacturing and on strengthening India's lead in services.

The key reason for the failure of labour-intensive manufacturing to flourish in India is the stringency of labour laws. As discussed in detail in Bhagwati and Panagariya (2012), labour laws become progressively more stringent as a firm's size increases above just six employees. At 100 workers, a manufacturing firm effectively loses the right to lay off workers under any circumstances, including bankruptcy. The law requires the firm to seek permission from the local state labour department to sack or make employees redundant, and the labour department almost never grants permission.

The result of these inhibiting labour laws has been that, in contrast to China, where large and medium firms dominate the employment scene, such firms are a rarity in India. In India, it is small firms that dominate the employment scene. In an important paper, Hasan and Jandoc (2012) compared the firm size distributions of India and China in the apparel industry. They found that these distributions are diametrically opposed to one another. In India, firms with 10 or fewer workers accounted for 87.4% of apparel employment in 2005. In China, 87.7% of apparel workers were employed in firms with 50 or more workers. Large firms in India typically dominate in highly capital-intensive sectors such as auto manufacture and auto parts, where labour costs are often a small part of the total costs.

Hasan and Jandoc also show that when large firms do exist in the labour-intensive sectors, they locate themselves with greater preponderance in states with less stringent labour laws. Similar cross-state differences do not arise in capital-intensive manufacturing sectors, nor do other factors, such as the availability of infrastructure in the states, produce similar cross-state differences. These factors point to an acute need for labour law reforms that would pave the way for the emergence of medium and large firms in labour-intensive sectors on a substantial scale. The small firms that currently dominate the scene simply do not have the incentive to exploit the vast world markets and are therefore unable to produce large numbers of well-paid jobs.

Reforms are, of course, required in several other areas as well. Land markets remain highly distorted, for instance. One particular area related to land in which reform is necessary and relatively straightforward is land acquisition. The current law governing land acquisition dates back to the 19th century and needs to be replaced by a more modern law that allows entrepreneurs to buy land freely from the current owners, including farm land, at competitive prices. There is also an acute

need for proper ownership titles to land and other property.

India also needs to build infrastructure in a major way. The past decade has seen considerable slowdown in the building of roads and this process must be reinvigorated. Likewise, with one-third of Indian households still without electricity, it is essential to reform the electricity sector. Electricity subsidies, which have left the distribution companies effectively bankrupt and discouraged electricity generation, need to be ended. Electricity tariffs need to be rationalised so that the industry is not charged punishing electricity prices to cross-subsidise other consumers.

India also needs to return forcefully to the opening of the external sector. Trade liberalisation has come to a virtual standstill under the current United Progressive Alliance, which originally came to power in May 2004. The outgoing government had dropped the top industrial tariff (with some exceptions) to 10% and it remains at that level to date. India will benefit from further liberalisation in all sectors: industry, agriculture and services. Liberalisation in agriculture has scarcely begun, so the scope for reform in this area is enormous. Likewise, in services, the multi-brand retail trade needs to be opened up to foreign investors. Large retailers from around the world can play a major role in modernising this sector. They can help build the supply chains both from retail to manufacturers and from manufacturers to the export markets.

To strengthen the services leg of the economy, higher education also needs urgent attention. The gross enrollment ratio in higher education at below 14% is a solid ten percentage points behind China. Shortages of qualified skilled workers may eventually slow the expansion of the information technology industry. Already, IT has been experiencing the fastest rise in wages over the last several years. In view of the need for an increased enrollment ratio, as well as to accommodate its burgeoning young population, India needs many hundreds of new universities. Given the fiscal constraints, the ability of the public sector to undertake the necessary expansion is limited. Therefore the conditions under which new private universities could be set up need to be liberalised as well. The current system, which requires legislation by Parliament or a state legislative assembly, is extremely cumbersome. What is needed, instead, is a set of administrative procedures that allow new universities to be opened up quickly and efficiently. India also needs to make the policy environment friendlier to virtual universities that can bring higher education to vast numbers of individuals at low cost.

A final important area in which India needs reforms is that of redistribution policies. The government has so far resisted even experimenting with direct cash transfers as the redistribution instrument. Instead, it has insisted on creating large supply chains in food distribution and the provision of education and health. Unfortunately, these supply chains have been hampered by corruption and huge inefficiencies that the government has been unable to keep in check. As a result, potential recipients of the service have steadily exited public supply, even when provided at highly subsidised rates. Cash transfers to the poor and allowing them to choose between public and private providers can considerably alleviate this problem. It will also empower the poor rather than providers. Under the current system, the poor are at the mercy of public sector providers. But once they hold the cash, they will be in a position to go to the provider of their choice, forcing the public provider either to improve efficiency or lose its business.

Concluding Remarks

Recent policy paralysis and the subsequent decline in the growth rate below 6% during the first two quarters of 2012 has driven home the lesson that government complacency is extremely costly. The 8–9% growth that the reforms of the Rao and Vajpayee administrations made possible over the last decade cannot be taken for granted. Continued reforms are required not only to sustain and accelerate the growth that has already been achieved, but also to prevent backsliding and an economic slowdown. The long-term reform agenda must address the continuing distortions in the factor markets. The creation of well-paid

jobs requires a policy environment that is friendly to large firms. In turn, this requires major reforms in labour laws that encourage rather than deter entrepreneurs from opting for labour-intensive technologies and labour-intensive sectors. Without such reform employment will be concentrated in tiny firms operating in the informal sector.

India needs to walk on two legs – manufacturing and services – and to that end needs to maintain momentum in services. Fulfilling this objective requires major reforms in higher education that help improve both the quality and quantity of skilled workers. It must invest in infrastructure, address land market distortions and liberalise trade. Evidence shows that economic growth in India has led to substantial reduction in poverty. Therefore, future growth will not only help eliminate the poverty that remains but also turn India a major global player. If the growth rate achieved during 2003–04 to 2010–11 is sustained, which is entirely feasible in view of the high investment rate and the competitiveness of the economy, India will become the third largest economy in the world by 2025. With its rising population of the young, it will also become a large supplier of the global workforce. The prospects for India to regain some of its lost glory have, thus, never been brighter.

See Also

- ▶ Caste System
- ▶ Financial Liberalization
- ▶ India, Economics in
- ▶ Indian Economic Development
- ▶ Poverty Lines

Bibliography

Bhagwati, J., and A. Panagariya. 2012. *India's tryst with destiny: Debunking myths that undermine progress and addressing new challenges*. New Delhi: HarperCollins.

Datt, G. 1998. Poverty in India and Indian states: An update. *International Food Policy Research Institute, Discussion Paper No. 47*, July.

Dehejia, R., and A. Panagariya. 2016. Services growth in India: A look inside the black box. In *Reforms and economic transformation in India*, ed. J. Bhagwati and A. Panagariya. New York: Oxford University Press.

Gupta, P., and A. Panagariya. 2012. Economic reforms and election outcomes. In *India's reforms: How they produced inclusive growth*, ed. J. Bhagwati and A. Panagariya. New York: Oxford University Press.

Hasan, R., and K.R.L. Jandoc. 2012. Labor regulations and the firm size distribution in Indian manufacturing. In *India's reforms: How they produced inclusive growth*, ed. J. Bhagwati and A. Panagariya. New York: Oxford University Press.

Hnatkovska, V., A. Lahiri, and S.B. Paul. 2012a. Castes and labor mobility. *American Economic Journal: Applied Economics* 4(2).

Hnatkovska, V., A. Lahiri, and S.B. Paul. 2012b. Breaking the caste barrier: Intergenerational mobility in India. *Journal of Human Resources*, forthcoming.

Krishna, P., and G. Sethupathy. 2012. Trade and inequality in India. In *India's reforms: How they produced inclusive growth*, ed. J. Bhagwati and A. Panagariya. New York: Oxford University Press.

Mukherji, J., and T. Ogawa. 2012. Will India be the first BRIC fallen angel? Available from: http://www.standardandpoors.com/spf/upload/Ratings_US/IndiaFirstBRICFallenAngel080612.pdf. Accessed 5 Sept 2012.

Mukim, M., and A. Panagariya. 2012. *A comprehensive look at poverty measures in India*. Program on Indian Economic Policies, Columbia University, forthcoming.

Panagariya, A. 2008. *India: The emerging giant*. New York: Oxford University Press.

Indicative Planning

Klaus Nielsen

Abstract

Indicative planning aims to coordinate private and public investment and output plans through forecasts or targets. Compliance is voluntary. The underlying logic is that the plan can supply economically valuable information which, as a public good, the market mechanism cannot disseminate efficiently. It may be perceived as a substitute for non-existing forward markets. However, indicative planning takes into account only endogenous market uncertainty, not exogenous uncertainty (technology, foreign trade and so on). Indicative planning has been most consistently and continuously implemented in France and Japan

but has been used in many other countries, although decreasingly so since the 1970s.

Keywords

Austrian economics; Bounded rationality; Forecasting; Forward markets; General equilibrium; Imperfect information; Indicative planning; Planning; Rational expectations; Uncertainty

JEL Classifications

D0

Indicative planning is a means of improving the performance of an economy through the elaboration of a set of consistent numerical forecasts or targets for the economic future. The aim is to coordinate private and public sector investment and output plans through the provision of economically valuable information. As distinct from directive central planning, as practised in the Soviet Union from the late 1920s, it is planning without compulsion. Compliance is purely voluntary. It is based on the idea that, if the plan is appropriately constructed, it will indicate an optimal path for the economy, which would then be spontaneously followed by the economic actors, without the need for compulsion. Decision-making is formally fully decentralized, but some versions of indicative planning include consultation with major private actors and the concertation of private investment plans. Furthermore, compliance is encouraged and facilitated by persuasion and cognitive framing and is sometimes supported by incentives. In addition, state-controlled investment funds may be guided into favoured projects in accordance with the plan. Furthermore, public sector commitment to implement planned public investment and output targets may constitute an element of certainty that facilitates the intended voluntary compliance.

The best-known examples of indicative planning are the plans elaborated by the French Commissariat Général du Plan and the Japanese Planning Agency since the Second World War. After the Second World War several European countries, such as the Netherlands, developed some sort of indicative planning, often linked to the building of multi-sector econometric models of the economy. Indicative planning was widely practised in developing countries during the post-war period until the 1980s (Belassa 1990). After the collapse of Communism, indicative planning was briefly adopted in Poland, and is still being used in some of the former republics of the Soviet Union. In 1965 an indicative National Plan was implemented in the United Kingdom, but was abandoned after a year as an effect of a balance of payment crisis. Today (in 2007), the European Union is involved in soft coordination activities that have some resemblance to indicative planning.

The presence of imperfect information is a market failure, and indicative planning can be seen as an attempt to bridge the information gap. The underlying logic is that the plan can supply economically valuable information which, as a public good, the market mechanism does not disseminate efficiently. Indicative planning makes it possible to overcome the problems that arise from the economic actors' ignorance of the intentions of the other actors. The collective market research involved in indicative planning should, in principle, make it possible to anticipate potential overcapacity and shortage and to avoid states of disequilibrium with unfulfilled expectations. If every economic actor informs the planners about their prospective demand and supply intentions for the forthcoming plan period, this information could be aggregated into an indicative plan and appropriate adaptations could be made by the economic actors.

The indicative plan may be perceived as a substitute for non-existing forward markets, or as a calculated general equilibrium representing an optimal allocation of resources that it would be in everybody's interest to implement on condition that the plan was correctly worked out. J.L. Meade (1971) demonstrates that the optimality features of the welfare-maximizing general equilibrium model can be obtained even if a full set of forward markets does not exist, provided that the economic agents make honest non-binding declarations about intended actions for any future date.

Based on this information, equilibrium prices and quantities could be calculated and the

forecasts of the indicative plan would necessarily be realized, since they correspond to optimal behaviour by market agents.

However, the assumption that agents declare their true intentions contradicts the assumption of rational behaviour if individual agents are large enough to influence prices that provide them with an incentive not to reveal their true preferences.

Furthermore, indicative planning is capable of taking into account only endogenous market uncertainty, and works only in a closed economy. Environmental, or exogenous, uncertainty (including changes in technology and foreign trade) is ignored. In theory, the indicative plan may operate with as many future paths as there are possible scenarios for the exogenous environment. However, this procedure for transformation of uncertainty to risk is hardly of any practical relevance, and it does not recognize the existence of genuine uncertainty that makes it impossible to elaborate appropriate scenarios, even in theory.

Economic internationalization and technological change have the effect that the overwhelming source of uncertainty has become exogenous, which has made the forecasting exercises of indicative planning increasingly difficult and ultimately useless. As a result, indicative planning has been widely abandoned or its ambitions have been significantly curtailed. France is the major example of a continuous commitment to indicative planning. Until 2006, planning documents covering successive 5-year planning periods were elaborated by the Commissariat Général du Plan. However, from the early 1970s and onwards, the plans became less ambitious and less influential. Targets and concertation were abandoned. The plans became internal governmental strategic documents that were, from 1993, no longer presented to Parliament. From 2006, indicative planning was formally abandoned, and the Commissariat Général du Plan was succeeded by a new Centre d'Analyse Stratégique.

It is fair to ask whether indicative planning, following its almost universal decline, is now devoid of contemporary relevance, if it ever had some, and has become a phenomenon of merely historical interest. Is indicative planning irrelevant, even in its less comprehensive and more pragmatic version that stresses the virtues of its contribution to develop shared expectations, or 'a common view of the future?' If the economic agents are seen as capable of developing rational expectations there is surely no role for indicative planning. In this view, attempts to influence expectations are ineffective and wasteful. From the point of view of Austrian economics, collective forecasting is even worse; it is not only ineffective but harmful. Indicative planning can be misleading, which may lead to too many eggs being put into one wrong basket. The plurality of information in a world of decentralized decision-making with no public attempts to influence expectations is seen as preferable by far.

However, from a more pragmatic point of view, it is exactly the role of indicative planning in forming common expectations concerning macroeconomic development trends that may contribute, not to the achievement of the nirvana of an optimal growth path, but rather to an improved state of disequilibrium (Holmes 1987). If optimal equilibrium is seen to be of little practical relevance as a result of widespread genuine uncertainty and the bounded rationality of economic agents, pragmatic means to improve the situation are important, although these may not in any way be seen as leading to a utopian state of optimal allocation of resources. At least three factors make indicative planning in the form of macroeconomic forecasts highly valuable in this context: (*a*) the public good character of the collected information, (*b*) the economies of scale of information processing, and (*c*) the fact that the government is no doubt a particularly well-informed actor in relation to macroeconomic developments.

See Also

- ▶ Forecasting
- ▶ Market Failure
- ▶ Planning
- ▶ Public Goods
- ▶ Uncertainty

Bibliography

Belassa, B. 1990. *Indicative planning in developing countries*, Policy research working paper no. 439. Washington, DC: World Bank.
Holmes, P. 1987. Indicative planning. In *The new Palgrave dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London: Macmillan.
Meade, J.E. 1971. *The theory of indicative planning*. Manchester: Manchester University Press.

Indicators

V. Zarnowitz

Types and Structure

Economic indicators, as a general category, are descriptive and anticipatory data used as tools for the analysis of business conditions and forecasting. There are potentially as many subsets of indicators in this sense as there are different targets at which they can be directed. For example, some indicators may relate to employment, others to inflation.

This brings to mind the uses of such time series as lagged explanatory variables in econometric models and regression equations. But there is a different, established meaning to what is often called the 'indicator approach'. This is a system of data and procedures designed to monitor, signal and confirm cyclical changes, especially turning points, in the economy at large. The series that serve this purpose are selected for being comprehensive and systematically related to business cycles and are known as *cyclical indicators*.

Business cycles are recurrent sequences of alternating phases of expansion and contraction that involve a great number of diverse economic processes. These movements are both sufficiently diffused and sufficiently synchronized to show up as distinct fluctuations in comprehensive series that measure production, employment, income and trade-aspects of aggregate economic activity. The end of each expansion is marked by a cluster of peaks in such series, the end of each contraction by a cluster of troughs. Analysts at the National Bureau of Economic Research (NBER) base the dating of business cycle peaks and troughs on the identification and analysis of such clusters, that is, the consensus of the corresponding turning points in the principal *coincident indicators*. This is done because (1) the co-movement of the indicators is itself an essential characteristic of the business cycle; (2) no single adequate measure of aggregate economic activity is available in a consistent form for a long historical period; and (3) economic statistics generally are subject to error, so that the evidence from a number of independently compiled indicators tends to be more reliable than the evidence from any individual series. The NBER reference chronologies of business cycle peaks and troughs (Burns and Mitchell 1946, ch. 4; Moore 1961, chs 5 and 6; Zarnowitz and Moore 1977, 1981) are widely used in academic as well as current business research.

The specific cycles observed across a wide spectrum of variables differ greatly and in part systematically. Thus many economic time series called the *leading indicators* tend to reach their turning points *before* the corresponding business cycle turns. There are also many series that tend to reach their turning points *after* the peaks and troughs in the business cycle, and they are the *lagging indicators*. The leading series represent largely flow and price variables that are highly sensitive to the overall cyclical influences but also to shorter random disturbances; hence they show large cyclical rises and declines but also high volatility. Coincident series have generally smaller cyclical movements and are at the same time much smoother. Lagging indicators include some massive stock variables which have modest cyclical functions yet are extremely smooth.

Most indicators display, in addition to the cyclical fluctuations that dominate the developments over spans of several years, trends that prevail across decades and reflect largely economic growth and, for nominal variables, inflation. Seasonal variations are likewise widespread but these stable or evolving patterns of intra-year change show much diversity, hence are often weakened by

aggregation, unlike the longer movements which are in general positively cross-correlated. It is a common practice to assume that the seasonal movements are exogeneous and separable, and cyclical indicators are used predominantly in 'seasonally adjusted' form. The reason is to show the trends and cycles in monthly or quarterly data more clearly, but control against significant errors from faulty seasonal adjustments is also necessary, and often neglected.

The indicators, then, are viewed as composites of trends, cyclical and 'irregular' movements. The latter are generally small and of stable random appearance, apart from occasional outliers due to some particular disturbances such as major strikes or unseasonable weather. The 'classical' decomposition approach does not rule out some interactions among the component movements. What is alien to the indicator analysis, however, is the more recent notion that the trends and cycles themselves are purely stochastic phenomena, essentially random walks or results of the cumulation of random changes.

In a growing economy business expansions must be on the average larger than contractions in terms of output, employment, etc., and they are also likely to be longer. The individual cycles and their phases, however, vary greatly in duration and amplitude. These differences are systematically related to the scope or diffusion of the cyclical movements among different units of observation (e.g. activities, regions, industries). Vigorous expansions are generally more widespread than weak expansions; severe contractions are more widespread then mild contractions. But the timing sequences and amplitude differences among the indicators are observed during long and short, strong and weak cycles.

Diffusion indexes are time series showing the percentage of items in a given population that are rising over a specified unit period. Information about the direction of the change can often be obtained much more readily than information about the size of the change, hence surveys designed to produce timely diffusion measures on actual or expected sales, prices, profits, etc., are popular in many countries. Moreover, diffusion indices are correlated with rates of change in the corresponding aggregates and tend to lead the levels of these aggregates.

Significance in Business Cycle Theories

The indicators in current use play important roles in many areas viewed as critical in business cycle theories. This is illustrated by the summary in Table 1, based on a long series of studies (for references, see Zarnowitz 1972, 1985; Moore 1983, pp. 347–351).

Indicators, Table 1

Theories or models	Some of the main factors	Evidence from time series
Accelerator-multiplier models; hypotheses on autonomous investment, innovations, and gestation lags	Interaction between investment, final demand and savings	Large cyclical movements in business investment commitments (orders, contracts) lead total output and employment; smaller movements in investment realizations (shipments, outlays) coincide or lag
Inventory investment models	Stock adjustments in response to sales changes and their effects on production	Inventory investment tends to lead; its declines during mild recessions are large relative to those in final sales
Old monetary over-investment and current monetarist theories	Changes in the supply of money, bank credit, interest rates, and the burden of private debt	Money and credit flows (rates of change) are highly sensitive, early leaders; velocity, market rates of interest, credit outstanding coincide or lag
Hypotheses of cost-price imbalances, volatility of prospective rates of return, and expectational errors	Changes in costs and prices, in the diffusion, margins, and totals of profits, and in business expectations	Profit variables and stock price indexes are sensitive early leaders. Unit labor costs lag

The literature on business cycles, though rich in ingenious hypotheses of varying plausibility and compatibility, produced no unified theory (Haberler 1964; Zarnowitz 1985). There is evidence in support of a number of different models that focus on period-specific or sector-specific aspects of the economy's motion. Monocausal theories may help explain some episodes but are invalidated by long experience. The regularities noted above are complementary in the interdependent economic system but some of them may be more important under certain temporarily prevailing conditions, others under different conditions. Thus, for business cycles analysis and forecasting, groups of leading, coincident and lagging indicators representing a whole set of these relationships are expected to outperform any individual indicators or subsets representing fewer regularities. This insight provides a general rationale for the line of research summarized below.

Selecting and Explaining the Principal Indicators

Cyclical indicators have been selected and analysed in a series of studies by the NBER and most recently by the Bureau of Economic Analysis (BEA) in the US Department of Commerce (Mitchell and Burns 1938; Moore 1950, 1961; Moore and Shiskin 1967; Zarnowitz and Boschan 1975a, b). The results include a cross-classification of over 100 series by several broad 'economic-process' groups (e.g. production and income, fixed capital investment, money and credit) and typical timing at business cycle peaks and troughs. The data are regularly presented in a monthly report of BEA, *Business Conditions Digest* (BCD). A detailed weighting scheme was developed to score each of these series by seven major criteria: economic significance, statistical adquacy, consistency of cyclical timing, conformity to business expansions and contractions, smoothness, prompt availability or currency, and reliability of preliminary as compared with revised data. As far as possible, the assessments were based on statistical measures to ensure their consistency and replicability.

The information thus collected served as a basis for the construction of *composite indexes* of leading, coincident and lagging indicators. These indexes incorporate the best-scoring series from the different economic-process categories and combine those with similar cyclical timing, using their overall performance scores as weights. The series are all monthly; all but a few, as noted below, represent real rather than nominal variables.

The coincident index comprises non-farm employment, industrial production, real personal income less transfer payments, and real manufacturing and trade sales. Repeated tests showed this index to have a better record of conformity, timing and currency than alternative indexes including real GNP and the unemployment rate.

There are good reasons to expect the sequences of the leading, coincident and lagging indexes to persist, as indeed they do. Several of the component leaders represent early stages of production and investment processes – commitments that precede the later stages of outlays, construction put in place and deliveries. This subset includes new business formation, contracts and orders for plant and equipment, new orders for consumer goods and materials, and permits for new housing.

The timing relations depend not only on technology but also on the state of the economy. Thus delivery periods get progressively longer just before and during recoveries and especially in booms when orders back up and strain the capacity to produce; and they get progressively shorter when an expansion slows down and a contraction develops. This explains the leads of vendor performance, percent of companies receiving slower deliveries and also, in part, the fact that the leads of the indicators tend to be considerably longer, but also more variable, at peaks than at troughs.

The change in manufacturing and trade inventories on hand and on order tends to turn before sales to which the desired level of the stocks is adjusted (a type of accelerator relationship). This series, a volatile mixture of intended and unintended investment, requires some smoothing. Total inventories move sluggishly; the ratio of inventories to sales is a component of the lagging index.

Sensitive prices of industrial materials are related to new orders, vendor performance and

inventory investment. The leading composite now includes the rate of change in an index of these prices but this is a very volatile series, even in somewhat smoothed form. In times of low inflation, the index itself (i.e. the level of such prices) would probably make a better indicator.

Another nominal indicator, the rate of change in business and consumer credit outstanding, leads because the new loans principally serve to finance investment in processes that are themselves leading (in inventories, housing and consumer durables; also in plant and equipment, where the loans are largely taken out early in the process). Here too, there are timing sequences that reflect stock-flow relationships: new increments lead, totals lag. The stock of commercial and industrial loans outstanding (deflated) is a component of the lagging index, and so is the ratio of consumer installment credit outstanding to personal income.

Compared with the overall credit flows, rates of growth in monetary aggregates show in general lower cyclical conformities and amplitudes and more random variations. They have historically led at business cycles turns by highly variable but mostly long intervals. The aggregates themselves are dominated by strong upward trends and show persistent declines only in cycles with severe contractions. However, a measure of 'real balances', the broadly defined money supply M2 deflated by a consumer price index, anticipated most of the recent business turns and is included in the current leading index. In late stages of expansion (contraction) money increased less (more) than prices.

The Standard & Poor's price index of 500 common stocks is included in the leading composite without adjustment for inflation. The market apparently tracks or anticipates well the movement of corporate earnings which is itself characterized by early timing. Money wages often rise less than prices in recoveries and more than prices late in expansion, while output per hour of labour fluctuates procyclically around a rising trend, generally with leads. Labour costs per unit of output, therefore, also move procyclically relative to their upward trends but with lags (they are a component of the lagging index). As a result of these tendencies connected with cyclical changes in sales and the rates of utilization of labour and capital, profit margins and totals swing widely in each cycle with sizeable leads.

Stock prices also tend to react inversely to changes in market interest rates. It is when an expansion (contraction) is well advanced and sufficiently strong that bank rates and bond yields tend to rise (decline) substantially, that is, interest rates generally lag. The average prime rate charged by banks is included in the lagging index.

Finally, there are the labour market indicators. Changes in hours are less binding than changes in the number employed, so the average workweek in manufacturing leads because it is altered early in response to uncertain signs of shifts in the demand for output. Initial claims for unemployment insurance lead the unemployment rate by short intervals. The average duration of unemployment lags the unemployment rate and is a component of the lagging index. These series, of course, show strong countercyclical movements, so they are used in inverted form.

Functions

When used collectively, the indicators provide over the course of business cycles a revolving flow of signals. Shallow and spotty declines in the leading series provide only weak and uncertain warnings; a run of several large declines increases a risk of a general and serious slow-down or recession. The latter may suggest some stabilizing policy actions which, if effective, could falsify the warning. The coincident indicators confirm or invalidate the expectations based on the behaviour of the leaders and any related policy decisions.

The lagging indicators provide further checks on the previously derived inferences, in particular on any early designation of the timing of a business cycle turn. Moreover, they also act as predictors. The turning points in the lagging index systematically precede the opposite turns in the leading index. Unit labour costs, interest rates, outstanding debt, and inventories measure or reflect the costs of doing business. For this reason, these series, when inverted, show very long leads.

For example, declines in inventories and interest rates during a recession pave the way for an upturn in new orders and then output of materials and finished goods (Zarnowitz and Boschan 1975b; Moore 1983, ch. 23).

Critique and Evidence

Enough has been said above on the reasons for the observed behaviour of indicators and their links to business cycle theories to weaken if not disprove the charge of 'measurement without theory'. If the reasons are simple so much the better. Macroeconomic forecasting, which the indicator system is designed to aid, must be essentially consistent with the ascertained regularities of business fluctuations, however difficult it may be to reconcile these 'stylized facts' with the preconceptions of general equilibrium theory.

The real problems with the indicators are mainly practical. Large amounts of random noise, large revisions of originally published figures, and short lead times (which occur mostly at troughs of short recessions) detract from the usefulness of some leading series. Those irregular variations and data errors in its components that are independent tend to cancel out in the leading index, which is therefore relatively smooth. This reduces but does not eliminate the problem of extra turns or false warnings. The index signalled each of the eight recessions but also each of the four major slow-downs (phases of below-average but still positive growth) in 1948–85. In sum, the leading indicators predict best the 'growth cycles', that is, fluctuations in trend-adjusted aggregates of output, employment, etc. This was found to be true as well for Japan, Canada and the major countries of Western Europe (Moore 1983, chs 5 and 6). A sequential signalling system designed to safeguard against false signals and discriminate in a timely fashion between recessions and slow-downs has been devised and tested with promising results (Zarnowitz and Moore 1982).

Forecasting with leading indicators has a long history of applications, elaborations, and revisions occasioned by new data and research findings, and changes in the workings of the economy (Burns 1950; Moore 1983, ch. 24). Repeated tests were made of both the turning-point predictions and forecasts of series such as real GNP and industrial production (Hymans 1973; Neftci 1979; Auerbach 1982). Tests have also been made by duplicating the US indicator test using data for other countries (Klein and Moore 1985). The most demanding, correctly performed tests produced generally positive results (see Auerbach 1982; Moore 1983, chs 24 and 25).

See Also

- ▶ Business Cycles
- ▶ Demand Management
- ▶ Stabilization Policy

Bibliography

Auerbach, A. 1982. The index of leading economic indicators: 'Measurement without theory' thirty-five years later. *Review of Economics and Statistics* 64(4): 589–595.

Burns, A.F. 1950. *New facts on business cycles*. New York: National Bureau of Economic Research.

Burns, A.F., and W.C. Mitchell. 1946. *Measuring business cycles*. New York: Columbia University Press for the National Bureau of Economic Research.

Haberler, G. 1964. *Prosperity and depression*, New ed. Cambridge, MA: Harvard University Press.

Hymans, S. 1973. *Brookings papers on economic activity*, vol. 2. Washington, DC: Bookings Institution.

Klein, P.A., and G.H. Moore. 1985. *Monitoring business cycles in market-oriented countries*. Cambridge, MA: Ballinger for the National Bureau of Economic Research.

Mitchell, W.C., and A.F. Burns. 1938. *Statistical indicators of cyclical revivals*. Bulletin 89. New York: National Bureau of Economic Research.

Moore, G.H. 1950. *Statistical indicators of cyclical revivals and recessions*. Occasional paper no. 31. New York: National Bureau of Economic Research.

Moore, G.H. 1961. *Business cycles indicators*. New York: National Bureau of Economic Research.

Moore, G.H. 1983. *Business cycles, inflation, and forecasting*, 2nd ed. Cambridge, MA: Ballinger for the National Bureau of Economic Research.

Moore, G.H., and J. Shiskin. 1967. *Indicators of business expansions and contractions*. New York: National Bureau of Economic Research.

Neftci, S. 1979. Leading-lag relations, exogeneity, and prediction of economic time series. *Econometrica* 47(1): 101–113.
Zarnowitz, V. (ed.). 1972. *The business cycle today*. New York: National Bureau of Economic Research.
Zarnowitz, V. 1985. Recent work on business cycles in historical perspective: A review of theories and evidence. *Journal of Economic Literature* 23(2): 523–280.
Zarnowitz, V., and C. Boschan. 1975a. Cyclical indicators: An evaluation and new leading index. *Business Conditions Digest*: v–xxii.
Zarnowitz, V., and C. Boschan. 1975b. New composite indexes of coincident and lagging indicators. *Business Conditions Digest*: v–xxiv.
Zarnowitz, V., and G.H. Moore. 1977. The recession and recovery of 1973–1976. *Explorations in Economic Research* 4(4): 471–557.
Zarnowitz, V., and G.H. Moore. 1981. The timing and severity of the 1980 recession. *NBER Reporter*, 19–21.
Zarnowitz, V., and G.H. Moore. 1982. Sequential signals of recession and recovery. *Journal of Business* 55(1): 57–85.

Indifference, Law Of

F. Y. Edgeworth

A designation applied by Jevons to the following fundamental proposition: 'In the same open market, at any one moment, there cannot be two prices for the same kind of article.'

This proposition, which is at the foundation of a large part of economic science, itself rests on certain ulterior grounds: namely, certain conditions of a perfect market. One is that monopolies should not exist, or at least should not exert that power in virtue of which a proprietor of a theatre, in Germany for instance, can make a different charge for the admission of soldiers and civilians, of men and women. The indivisibility of the articles dealt in appears to be another circumstance which may counteract the law of indifference in some kinds of market, where price is not regulated by cost of production.

[Jevons, *Theory of exchange,* 2nd ed, p. 99 (statement of the law). Walker, *Political economy,* art. 132 (a restatement). Mill, *Political economy,* bk. ii. ch. iv. § 3 (imperfections of actual markets). Edgeworth, *Mathematical psychics,* pp. 19, 46 (possible exceptions to the law of indifference).]

Bibliography

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.
Jevons, W.S. 1875. *Money, the mechanism of exchange*. London: D. Appleton and co.
Mill, J.S. 1848. *Principles of political economy*. London: J.W. Parker.
Walker, F.A. 1886. *A brief textbook of political economy*. London: Macmillan.

Indirect Inference

Anthony A. Smith, Jr.

Abstract

Indirect inference is a simulation-based method for estimating the parameters of economic models. Its hallmark is the use of an auxiliary model to capture aspects of the data upon which to base the estimation. The parameters of the auxiliary model can be estimated using either the observed data or data simulated from the economic model. Indirect inference chooses the parameters of the economic model so that these two estimates of the parameters of the auxiliary model are as close as possible. The auxiliary model need not be correctly specified; when it is, indirect inference is equivalent to maximum likelihood.

Keywords

Auxiliary models; Bayesian inference; Criterion functions; Discrete-choice models; Dynamic stochastic general equilibrium (DSGE) models; Estimation; Lagrange multipliers; Likelihood; Likelihood ratios; Linear probability models; Maximum likelihood; Models; Probability density functions;

Reduced-form models; Seminonparametric (SNP) models; Simulated moments estimation; Simultaneous equations; Vector autoregressions; Wald test

JEL Classifications
C52

Indirect inference is a simulation-based method for estimating, or making inferences about, the parameters of economic models. It is most useful in estimating models for which the likelihood function (or any other criterion function that might form the basis of estimation) is analytically intractable or too difficult to evaluate. Such models abound in modern economic analysis and include nonlinear dynamic models, models with latent (or unobserved) variables, and models with missing or incomplete data.

Like other simulation-based methods, indirect inference requires only that it be possible to simulate data from the economic model for different values of its parameters. Unlike other simulation-based methods, indirect inference uses an approximate, or auxiliary, model to form a criterion function. The auxiliary model does not need to be an accurate description of the data generating process. Instead, the auxiliary model serves as a window through which to view both the actual, observed data and the simulated data generated by the economic model: it selects aspects of the data upon which to focus the analysis.

The goal of indirect inference is to choose the parameters of the economic model so that the observed data and the simulated data look the same from the vantage point of the chosen window (or auxiliary model). In practice, the auxiliary model is itself characterized by a set of parameters. These parameters can themselves be estimated using either the observed data or the simulated data. Indirect inference chooses the parameters of the underlying economic model so that these two sets of estimates of the parameters of the auxiliary model are as close as possible.

A Formal Definition

To put these ideas in concrete form, suppose that the economic model takes the form:

$$y_t = G(y_{t-1}, x_t, u_t; \beta), t = 1, 2, \ldots, T, \qquad (1)$$

where $\{x_t\}_{t=1}^{T}$ is a sequence of observed exogenous variables, $\{y_t\}_{t=1}^{T}$ is a sequence of observed endogenous variables, and $\{u_t\}_{t=1}^{T}$ is a sequence of unobserved random errors. Assume that the initial value y_0 is known and that the random errors are independent and identically distributed (i.i.d.) with a known probability distribution F. Equation (1) determines, in effect, a probability density function for y_t conditional on y_{t-1} and x_t. Indirect inference does not require analytical tractability of this density, relying instead on numerical simulation of the economic model. This is not the most general model that indirect inference can accommodate – indirect inference can be used to estimate virtually any model from which it is possible to simulate data – but it is a useful starting point for understanding the principles underlying indirect inference. The econometrician seeks to use the observed data to estimate the k-dimensional parameter vector β.

The auxiliary model, in turn, is defined by a conditional probability density function, $f(y_t \mid y_{t-1}, x_t, 0)$, which depends on a p-dimensional parameter vector θ. In a typical application of indirect inference, this density has a convenient analytical expression. The number of parameters in the auxiliary model must be at least as large as the number of parameters in the economic model (that is, $p \geq k$).

The auxiliary model is, in general, incorrectly specified: that is, the density f need not describe accurately the conditional distribution of y_t determined by Eq. (1). Nonetheless, the parameters of the auxiliary model can be estimated using the observed data by maximizing the log of the likelihood function defined by f:

$$\hat{\theta} = \arg\max_{\theta} \sum_{t=1}^{T} \log f(y_t \mid y_{t-1}, x_t, \theta).$$

The estimated parameter vector $\hat{\theta}$ serves as a set of 'statistics' that capture, or summarize, certain features of the observed data; indirect inference chooses the parameters of the economic model to reproduce this set of statistics as closely as possible.

The parameters of the auxiliary model can also be estimated using simulated data generated by the economic model. First, using a random number generator, draw a sequence of random errors $\{\tilde{u}_t^m\}_{t=1}^T$ from the distribution F. Typically, indirect inference uses M such sequences, so the superscript m indicates the number of the simulation. These sequences are drawn only once and then held fixed throughout the estimation procedure. Second, pick a parameter vector β and then iterate on Eq. (1), using the observed exogenous variables and the simulated random errors, to generate a simulated sequence of endogenous variables: $\{\tilde{y}_t^m(\beta)\}_{t=1}^T$, where the dependence of this simulated sequence on β is made explicit. Third and finally, maximize the average of the log of the likelihood across the M simulations to obtain:

$$\hat{\theta}(\beta) = \arg\max_{\theta} \sum_{m=1}^{M} \sum_{t=1}^{T} \log f\left(\tilde{y}_t^m(\beta) \middle| \tilde{y}_{t-1}^m(\beta), x_t, \theta\right).$$

The central idea of indirect inference is to choose β so that $\hat{\theta}(\beta)$ and $\hat{\theta}$ are as close as possible. When the economic model is exactly identified (that is, when $p = k$), it is, in general, possible to choose β so that the economic model reproduces exactly the estimated parameters of the auxiliary model. Typically, though, the economic model is over-identified (that is, $p > k$): in this case, it is necessary to choose a metric for measuring the distance between $\hat{\theta}$ and $\hat{\theta}(\beta)$; indirect inference then picks β to minimize this distance.

As the observed sample size T grows large (with M held fixed), the estimated parameter vector in the simulated data, $\hat{\theta}(\beta)$, converges to a so-called 'pseudo-true value' that depends on β; call it $h(\beta)$. The function h is sometimes called the binding function: it maps the parameters of the economic model into the parameters of the auxiliary model. Similarly, the estimated parameter vector in the observed data, $\hat{\theta}$, converges to a pseudo-true value θ_0. In the limit as T grows large, then, indirect inference chooses β to satisfy the equation $\theta_0 = h(\beta)$. Under the assumption that the observed data is generated by the economic model for a particular value, β_0, of its parameter vector, the value of β that satisfies this equation is precisely β_0. This heuristic argument explains why indirect inference generates consistent estimates of the parameters of the economic model.

Three Examples

Example 1: A Simple System of Simultaneous Equations

The first example is drawn from the classical literature on simultaneous equations to which indirect inference is, in many ways, a close cousin. Consider a simple macroeconomic model, adapted from Johnston (1984), with two simultaneous equations: $C_t = \beta Y_t + u_t$ and $Y_t = C_t + X_t$. In this model, consumption expenditure in period t, C_t, and output (or income) in period t, Y_t, are endogenous, whereas nonconsumption expenditure in period t, X_t, is exogenous. Assume that the random error u_t is i.i.d. and normally distributed with mean zero and a known variance; the only unknown parameter, then, is β.

There are many ways to estimate β without using indirect inference, but this example is useful for illustrating how indirect inference works. To wit, suppose that the auxiliary model specifies that C_t is normally distributed with conditional mean θX_t and a fixed variance. In this simple example, the binding function can be computed without using simulation: a little algebra reveals that $\theta = \beta/(1 - \beta) \equiv h(\beta)$. To estimate β, first use ordinary least squares (which is equivalent to maximum likelihood in this example) to obtain a consistent estimate, $\hat{\theta}$, of θ. Then evaluate the inverse of h at $\hat{\theta}$ to obtain a consistent estimate of β : $\hat{\beta} = \hat{\theta}/\left(1 + \hat{\theta}\right)$. This is precisely the indirect inference estimator of β. This estimator uses an indirect approach: it first estimates an auxiliary (or, in the language of simultaneous equations, a reduced-form) model whose parameters are

complicated functions of the parameters of the underlying economic model and then works backwards to recover estimates of these parameters.

Example 2: A General Equilibrium Model of the Macroeconomy

In this example, the economic model is a dynamic, stochastic, general equilibrium (DSGE) model of the macroeconomy (for a prototype, see Hansen 1985). Given choices for the parameters describing the economic environment, this class of models determines the evolution of aggregate macroeconomic time series such as output, consumption, and the capital stock. The law of motion for these variables implied by the economic model is, in general, nonlinear. In addition, some of the key variables in this law of motion (for example, the capital stock) are poorly measured or even unobserved. For these reasons, in these models it is often difficult to obtain a closedform expression for the likelihood function.

To surmount these obstacles, indirect inference can be used to obtain estimates of the parameters of the economic model. A natural choice for the auxiliary model is a vector autoregression (VAR) for the variables of interest. As an example, let y_t be a vector containing the values of output and consumption in period t (expressed as deviations from steady-state values) and let the VAR for y_t have one lag: $y_{t+1} = Ay_t + \varepsilon_{t+1}$, where the ε_ts are normally distributed, i.i.d. random variables with mean 0 and covariance matrix Σ.

In this example, the binding function maps the parameters of the economic model into the parameters A and Σ of the VAR. To obtain a simulated approximation to the binding function, pick a set of parameters for the economic model, compute the law of motion implied by this set of parameters, simulate data using this law of motion, and then use OLS to fit a VAR to the simulated data. Indirect inference chooses the parameters of the economic model so that the VAR parameters implied by the model are as close as possible to the VAR parameters estimated using observed macroeconomic time series. Smith (1993) illustrates the use of indirect inference to estimate DSGE models.

Example 3: A Discrete-Choice Model

In this example, the economic model describes the behaviour of a decision-maker who must choose one of several discrete alternatives. These models typically specify a random utility for each alternative; the decision-maker is assumed to pick the alternative with the highest utility. The random utilities are latent: the econometrician does not observe them, but does observe the decision-maker's choice. Except in special cases, evaluating the likelihood of the observed discrete choices requires the evaluation of high-dimensional integrals which do not have closed-form expressions.

To use indirect inference to estimate discrete-choice models, one possible choice for the auxiliary model is a linear probability model. In this case, the binding function maps the parameters describing the probability distribution of the latent random utilities into the parameters of the linear probability model. Indirect inference chooses the parameters of the economic model so that the estimated parameters of the linear probability model using the observed data are as close as possible to those obtained using the simulated data. Implementing indirect inference in discrete-choice models poses a potentially difficult computational problem because it requires the optimization of a non-smooth objective function. Keane and Smith (2003), who illustrate the use of indirect inference to estimate discrete-choice models, also suggest a way to smooth the objective surface.

Three Metrics

To implement indirect inference when the economic model is over-identified, it is necessary to choose a metric for measuring the distance between the auxiliary model parameters estimated using the observed data and the simulated data, respectively. There are three possibilities corresponding to the three classical hypothesis tests: Wald, likelihood ratio (LR), and Lagrange multiplier (LM).

In the Wald approach, the indirect inference estimator of the parameters of the economic model minimizes a quadratic form in the

difference between the two vectors of estimated parameters:

$$\hat{\beta}^{Wald} = \arg\min_{\beta} \left(\hat{\theta} - \hat{\theta}(\beta)\right)' W \left(\hat{\theta} - \hat{\theta}(\beta)\right),$$

where W is a positive definite 'weighting' matrix.

The LR approach to indirect inference forms a metric using the (approximate) likelihood function defined by the auxiliary model. In particular,

$$\hat{\beta}^{LR} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \log f\left(y_t | y_{t-1}, x_t, \hat{\theta}\right) - \sum_{t=1}^{T} \log f\left(y_t | y_{t-1}, x_t, \hat{\theta}(\beta)\right). \right).$$

By the definition of $\hat{\theta}$, the objective function on the right-hand side is non-negative, and its value approaches zero as $\hat{\theta}(\beta)$ approaches $\hat{\theta}$. The LR approach to indirect inference chooses β so as to make this value as close to zero as possible. Because the first term on the right-hand side does not depend on β, the LR approach can also be viewed as maximizing the approximate likelihood subject to the restrictions, summarized (for large T) by the binding function h, that the economic model imposes on the parameters of the auxiliary model.

Finally, the LM approach to indirect inference forms a metric using the derivative (or score) of the log of the likelihood function defined by the auxiliary model. In particular,

$$\hat{\beta}^{LM} = \arg\min_{\beta} S(\beta)' V S(\beta),$$

where

$$S(\beta) = \sum_{m-1}^{M} \sum_{t=1}^{T} \frac{\partial}{\partial \theta} \log f\left(\tilde{y}_t^m(\beta), x_t, \hat{\theta}\right)$$

and V is a positive definite matrix. By definition, $\hat{\theta}$ sets the score in the observed data to zero. The goal of the LM approach, then, is to choose β so that the (average) score in the simulated data, evaluated at $\hat{\theta}$, is as close to zero as possible.

For any number, M, of simulated data-sets, all three approaches deliver consistent and asymptotically normal estimates of β as T grows large. The use of simulation inflates asymptotic standard errors by the factor $(1 + M^{-1})^{1/2}$; for $M \geq 10$, this factor is negligible. When the economic model is exactly identified, all three approaches to indirect inference yield numerically identical estimates; in this case, they all choose β to solve $\hat{\theta}(\beta) = \hat{\theta}$.

When the economic model is over-identified, the minimized values of the three metrics are, in general, greater than zero. These minimized values can be used to test the hypothesis that the economic model is correctly specified: sufficiently large minimized values constitute evidence against the economic model.

If the weighting matrices W and V are chosen appropriately, then the Wald and LM approaches are asymptotically equivalent in the sense that they have the same asymptotic covariance matrix; by contrast, the LR approach, in general, has a larger asymptotic covariance matrix. If, however, the auxiliary model is correctly specified, then all three approaches are asymptotically equivalent not only to each other but also to maximum likelihood (for large M). Because maximum likelihood is asymptotically efficient (that is, its asymptotic covariance matrix is as small as possible), the LM approach is sometimes called the 'efficient method of moments' when the auxiliary model is close to being correctly specified; in such a case, this name could also be applied to the Wald approach.

When estimating the parameters of the auxiliary model is difficult or timeconsuming, the LM approach has an important computational advantage over the other two approaches. In particular, it does not require that the auxiliary model be estimated repeatedly for different values of the parameters of the economic model. To estimate continuous-time models of asset prices, for example, Gallant and Tauchen (2005) advocate using a semi-nonparametric (SNP) model as the auxiliary model. As the number of its parameters increases, an SNP model provides an arbitrarily accurate approximation to the data generating process, thereby permitting indirect inference to approach the asymptotic efficiency of maximum likelihood. For this class of auxiliary models, which are nonlinear and often have a large number of parameters, the LM approach is a computationally attractive way to implement indirect inference.

Concluding Remarks

Indirect inference is a simulation-based method for estimating the parameters of economic models. Like other simulation-based methods, such as simulated moments estimation (see, for example, Duffie and Singleton 1993), it requires little analytical tractability, relying instead on numerical simulation of the economic model. Unlike other methods, the 'moments' that guide the estimation of the parameters of the economic model are themselves the parameters of an auxiliary model. If the auxiliary model comes close to providing a correct statistical description of the economic model, then indirect inference comes close to matching the asymptotic efficiency of maximum likelihood. In many applications, however, the auxiliary model is chosen, not to provide a good statistical description of the economic model, but instead to select important features of the data upon which to focus the analysis.

There is a large literature on indirect inference, much of which is beyond the scope of this article. Gouriéroux and Monfort (1996) provide a useful survey of indirect inference. Indirect inference was first introduced by Smith (1990, 1993) and later extended in important ways by Gouriéroux et al. (1993) and Gallant and Tauchen (1996). Although indirect inference is a classical estimation method, Gallant and McCulloch (2004) show how ideas from indirect inference can be used to conduct Bayesian inference in models with intractable likelihood functions. There have been many interesting applications of indirect inference to the estimation of economic models, mainly in finance, macroeconomics, and labour economics. Because of its flexibility, indirect inference can be a useful way to estimate models in all areas of economics.

See Also

► Maximum Likelihood

Bibliography

Duffie, D., and K.J. Singleton. 1993. Simulated moments estimation of Markov models of asset prices. *Econometrica* 61: 929–952.

Gallant, A.R., and R. McCulloch. 2004. On the determination of general scientific models. Working paper. Fuqua School of Business/Duke University.

Gallant, A.R., and G. Tauchen. 1996. Which moments to match? *Econometric Theory* 12: 657–681.

Gallant, A.R., and G. Tauchen. 2005. Simulated score methods and indirect inference for continuous-time models. In *Handbook of financial econometrics*, ed. Y. Ait-Sahalia and L. Hansen. Amsterdam: North-Holland.

Gouriéroux, C., and A. Monfort. 1996. *Simulation-based econometric methods*. New York: Oxford University Press.

Gouriéroux, C., A. Monfort, and E. Renault. 1993. Indirect inference. *Journal of Applied Econometrics* 8: S85–S118.

Hansen, G.D. 1985. Indivisible labor and the business cycle. *Journal of Monetary Economics* 16: 402–417.

Johnston, J. 1984. *Econometric methods*. 3rd edn. New York: McGraw-Hill Book Company.

Keane, M., and A.A. Smith, Jr. 2003. Generalized indirect inference for discrete choice models. Working paper. Yale University.

Smith, A.A., Jr. 1990. Three essays on the solution and estimation of dynamic macroeconomic models. Doctoral dissertation. Duke University.

Smith, A.A. Jr. 1993. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics* 8: S63–S84.

Indirect Taxes

John Kay

It is conventional to describe direct taxes as taxes where the person legally liable to pay the tax is also the person whose income or welfare is reduced as a result of its imposition: while indirect taxes are those where liability can be shifted to someone else. This distinction is essentially an arbitrary one. All taxes can be shifted to some degree: only in exceptional circumstances can any agent shift a tax completely. In common usage, indirect taxes are those which are paid by retailers, wholesalers or manufacturers, but

believed to be shifted toward final consumers. In this entry indirect taxation is regarded as synonymous with commodity taxation.

There are three main categories of indirect tax. Excise duties fall on particular commodities, especially those goods which are traditionally subject to particularly heavy taxation, such as tobacco products and alcoholic drinks. Such taxes are often specific – charged per unit of the commodity concerned – but may be ad valorem – assessed as a percentage of the retail price. In many countries these specific tax rates have failed to keep pace with inflation, and this has tended to reduce the incidence of these taxes in relation to total revenue and to the price of the commodities concerned.

More broadly based indirect taxes are usually set at ad valorem rates. These may be single stage taxes, collected at wholesale or retail level. Alternatively, multi-stage taxes may be imposed at each part of the production process. Turnover or cascade taxes are of this kind, but the most common multi-stage tax is a value added tax. This has been obligatory for member states of the EEC since directives based on the Neumark Report of 1963, and is now used in around forty countries. This tax is payable at all stages of production, but recoverable by all business purchasers, who are however obliged to charge tax on their own output. The consequence is that net tax is payable only on sales to final consumers.

The value added tax is consistent with one important result from the theory of commodity taxation. This is that commodity taxes should be levied only on purchases of goods by final consumers, and not on intermediate transactions between producers (Diamond and Mirrlees 1971). The reason for this is that the distributional or other objectives of commodity taxation can in all cases be achieved equally well by the taxation of final commodities; the taxation of intermediate goods achieves no advantage in this but additionally distorts the choices of inputs made by producers. It therefore imposes the avoidable distortion of production inefficiency on top of the inevitable deadweight loss in consumption which is common to any system of commodity taxation.

Commodity taxes impose deadweight losses on consumers. These arise from the distortions of consumer choice which create costs over and above the tax revenue derived by governments. Traditionally these have been expressed in terms of consumer surplus triangles but are now more effectively expressed using the dual formulation of demand theory implied by the expenditure function (Diamond and McFadden 1974). This gives deadweight loss as

$$\mathrm{L} = \mathrm{E}(p^{+}t, u) - \mathrm{E}(p, u) - tx$$

where p is a vector of producer prices, t is the tax vector, x the purchased commodity vector and u the reference utility level for evaluation of the expenditure function.

The optimal structure of commodity taxation may be derived from the minimization of L, and this leads to two schools of thought on the appropriate structure of commodity taxes. By choosing a vector t to minimize L we derive the Ramsey (1927) rules for optimal commodity tax rates in the implicit form

$$\sum_j t_j \left(\frac{\partial x_i}{\partial t_j}\right)_u = \lambda x_i$$

for some λ increasing in tax revenue. This can be interpreted as requiring that the compensated demand for all goods should be reduced in the same proportion. If there is no net complementarity or substitutability, this condition reduces to

$$\frac{t_i}{p_i} = \lambda \frac{x_i}{p_i}\left(\frac{\partial x_i}{\partial t_i}\right)_u$$

which yields an inverse elasticity rule: tax should bear most heavily on commodities in inelastic demand.

A weakness of this analysis is that L is still more effectively minimized – indeed reduced to zero – by the imposition of a lump sum tax. While lump sum taxes varying across individuals and independent of their economic behaviour are generally reckoned to be impracticable, a uniform lump sum is feasible. The primary reasons for

rejecting a poll tax – concern for distributional effects among households with different tastes and endowments – are abstracted from in the Ramsey formulation. A more direct way of reaching the same conclusion is to observe that life is the most inelastically demanded commodity of all. It follows that any set of Ramsey taxes will always be dominated by a poll tax.

If, however, a distributional objective is introduced then any set of commodity taxes will generally be inferior to a tax related to the total income of the consumer (or his total consumption, since there is no difference at the present level of abstraction). Thus there is a role for commodity taxes only if they are related to other household characteristics which cannot be observed and taxed directly, such as household skill levels. The argument illustrates a general feature of recent optimal tax theory, which shows that efficient tax structures are often very sensitive to the assumptions made about the other policy instruments available.

These results direct attention towards a different tradition (see, for example Hotelling 1938), which favours uniformity of rates of commodity taxes, on the grounds that this leaves relative commodity prices equal to relative marginal costs. This would be appropriate if all commodities were taxable, but there is at least one important good – leisure – which cannot be subjected to taxation. This takes the problem of optimal commodity taxes into the realm of the second best and suggests relatively high rates of taxation on those goods which are complementary with leisure and lower rates on those which are substitutes for it.

There are other reasons for departures from uniformity. Merit goods (Musgrave 1959) are commodities, such as education, whose consumption is thought to have some value, either social or for the individual concerned, beyond his own personal assessment. This may be a reason for specially low rates of tax on particular commodities or, more commonly, for specially high excise taxes. Corrective taxes are also a means by which market outcomes can be induced to reflect the externalities – good or bad – which are associated with particular kinds of production or consumption.

With these exceptions, the theoretical arguments for extensive departures from a general principle of uniformity in commodity taxation do not seem strong. In the main, most objectives which governments seek through elaborately differentiated rate structures can be more effectively achieved in other ways. Since this uniformity has considerable administrative advantages, both analytical and practical considerations point in a similar direction. The widespread move throughout the world to broadly based value-added taxes as a primary instrument of indirect taxation reflects the application of these principles.

See Also

- ▶ Public Finance
- ▶ Tax Incidence
- ▶ Value-Added Tax

Bibliography

Atkinson, A.B.., and J.E. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw-Hill.

Diamond, P.A., and D. McFadden. 1974. Some uses of the expenditure function in public finance. *Journal of Public Economics* 3(1): 3–21.

Diamond, P.A., and Mirrlees, J.A. 1971. Optimal taxation and public production, Parts I and II. *American Economic Review* 61: 8–27 and 261–278.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6: 242–269.

Musgrave, R.A. 1959. *The theory of public finance*. New York: McGraw-Hill.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Indirect Utility Function

Peter Newman

JEL Classifications
D1

After many independent discoveries that were widely separated in time and space, the indirect utility function has in the last 35 years gradually

become a standard part of demand theory. Its first discovery was made as early as 1886 by Antonelli in Italy, who also derived what has come to be known as Roy's Identity (see Chipman's introduction to the translation of Antonelli (1886) in Chipman et al. 1971). Later contributions came from Konyus (1924, 1926) and Byushgens in Russia, from Hotelling (1932) and Court (1941, pp. 284–97) in the United States, from Roy (1942, 1947) and Ville (1946) in France, and from Wold (1943–4) and Malmquist (1953) in Sweden; a good brief history may be found in Diewert (1982, pp. 547–50).

But it was not until the early 1950s and the contributions of Houthakker (1951–2, 1960) that the indirect utility function became an integral part of the theory of consumer's behaviour. Indeed, the very names in standard use appear to be due to him, 'indirect utility function' in (1951–2, p. 157) and 'Roy's Identity' in (1960, p. 250).

Definition and Simple Properties

Suppose that the consumer has completely preordered preferences defined over the commodity space R^{n+} of non-negative bundles $x = (x_1, x_2, \ldots, x_n)$, that those preferences are representable by a real-valued utility function u, that he (or she) faces competitively determined positive money prices $(p_1, p_2, \ldots, p_n) = p$ for the n goods, and has exogenously determined monetary wealth $\omega > 0$. It is standard in demand theory to assume that the consumer chooses a bundle x^* by solving the optimization problem:

$$\text{Max } (p, \omega) \text{ Find } x \in R^{n+} \text{ to max } u(x) \text{ subject to } \langle p, x \rangle \leq \omega \quad (1)$$

where the notation $\langle \cdot, \cdot \rangle$ means the inner product of the two vectors concerned.

Assume that Max (p, ω) has a unique solution x^*, for which it suffices that preferences be monotonically increasing and strictly convex. Then the number

$$\tau^* = u(x^*) \quad (2)$$

is the *value* of Max(p, ω). This joint determination of solution and value once p and ω are known implies the existence of two functions of the price-wealth pair (p, ω), called respectively the *ordinary* (or Marshallian) *demand function* f: $R^{n++} \times R^{++} \to R^{n++}$, defined by

$$x^* = f(p, \omega) \quad (3)$$

and the *indirect utility function* v: $R^{n++} \times R^{++} \to R$, defined by

$$\tau^* = v(p, \omega) \quad (4)$$

Define the attainable (or budget) set $A(p, \omega)$ by

$$A(p, \omega) = \{x \in R^{n+} : \langle p, w \rangle \leq \omega\}$$

From (1) it follows that for any $\lambda > 0$, $A(\lambda p, \lambda\omega) = A(p, \omega)$, so that both f and v are positively homogeneous of degree zero in (p, ω). Next, if $(p^1 - p^2) \in R^{n+}$ and $p^1 \neq p^2$ then $A(p^1, \omega) \subset A(p^2, \omega)$, from which $v(p^1, \omega) \leq v(p^2, \omega)$; for similar reasons, $v(p, \cdot)$ is nondecreasing. It can be shown further that if u is continuous then so is v (see e.g. Varian 1984, pp. 121, 326–7).

A useful result is that $v(\cdot, \omega)$ is quasi-convex. To prove this let $p^t = tp^1 + (1 - t)p^2$, where $t \in [0, 1]$. Then for any $x \in A(p^t, \omega)$,

$$t\langle p^1, x \rangle + (1 - t)\langle p^2, x \rangle < \omega. \quad (5)$$

If $t = 0$, $x \in A(p^2, \omega)$, while if $t = 1$, $x \in A(p^1, \omega)$. Otherwise, suppose that x is in neither $A(p^1, \omega)$ nor $A(p^2, \omega)$. Then $t\langle p^1, x \rangle > t\omega$ and $(1 - t)\langle p^2, x \rangle > (1 - t)\omega$, which on addition yield a contradiction to (5). So x is in either $A(p^1, \omega)$ or $A(p^2, \omega)$. Hence $v(p^t, \omega)$, which is the sup of $u(\cdot)$ on $A(p^1, \omega)$, can be no larger than $\max[v(p^1, \omega), v(p^2, \omega)]$, which are themselves the sups of u$(\cdot)$ on $A(p^1, \omega)$ and $A(p^2, \omega)$, respectively. But the condition $v(p^t, \omega) \leq \max[v(p^1, \omega), v(p^2, \omega)]$ is the original definition of the quasi-convexity of the function $v(\cdot, \omega)$ (see Fenchel 1953, p. 117).

Relations Between the Ordinary Demand Functions and the Indirect Utility Function

For simplicity, the following assumptions are made: (a) x^* is a strictly positive vector. (b) Each function involved is as differentiable as required. (c) At any $x \in R^{n+}$ there is at least one commodity in which u is strictly increasing (this implies local non-satiation of preferences).

Suppose that at x^* the constraint (1) is 'slack', i.e. $\omega - \langle p, x^* \rangle = \delta > 0$. Let k be any good with property (c) at x^*, and define a new bundle x^1 by putting $x_i^i = x_i^*$ for $i \neq k$, and $x_k^1 = x_k^* + (\delta/p_k)$. Then by construction $\langle p, x^1 \rangle = \omega$, while from (c) $u(x^1) > u(x^*)$, contradicting the hypothesis that x^* solves max(p, ω). So

$$\langle p, x^* \rangle = \omega \tag{6}$$

Next, define $L: R^{n+} \times R^{n++} \rightarrow R$ by

$$L(x^1, p^1) = v(p^1, \langle p^1, x^1 \rangle) - u(x^1) \tag{7}$$

where x^1 and p^1 are arbitrary. From (2) and (4), for any x^1 the value $v(p^1, \langle p^1\ x^1 \rangle)$is the *maximized* level of utility when prices are p^1 and wealth $\langle p^1, x^1 \rangle > 0$. Hence, $L(\cdot, p^1)$ is positive semi-definite, i.e. $x^1 \in R^{n+}$ implies $L(x^1, p^1) \geq 0$ for any p^1. Putting $p^1 = p$, the actual prices, if x^* solves Max(p, ω) it follows from (2), (4) and (6) that

$$L(x^*, p) = v(p, \langle p, x^* \rangle) - u(x^*) = 0 \tag{8}$$

Hence x^* attains the infimum of L$(\cdot, p)$. So from (6), (8) and the Chain Rule,

$$\forall i = 1, 2 \ldots, n \quad v\omega(p, \omega) p_i = u_i(x^*) \tag{9}$$

From (c), $u_i(x^*) > 0$ for at least one i. Since $p_i > 0$ this implies the simple but important result

$$v\omega(p, \omega) > 0 \tag{10}$$

i.e. the marginal utility of wealth is positive.

From (2), (3) and (4) the equation

$$v(p, \omega) = (f(p, \omega))$$

is an identity in (p, ω). So differentiating each of the individual demand functions f_i with respect to (wrt) each p_j and ω yields,

$$\begin{gathered} \forall\ j = 1, 2, \ldots, n \quad v_j(p, \omega) = \sum u_i(x^*) f_{ij}(p, \omega) \\ v\omega(p, \omega) = \sum u_i(x^*) f_i \omega(p, \omega) \end{gathered} \tag{11}$$

From (6) and (3),

$$\langle p, f(p, \omega) \rangle = \omega$$

This is another identity in (p, ω), and differentiating it wrt each p_j and ω results in

$$\begin{gathered} \forall j = 1, 2, \ldots, n \ f_i(p, \omega) + \Sigma p_i f_{ij}(p, \omega) = 0 \\ \Sigma p_i f_{i\omega}(p, \omega) = 1 \end{gathered} \tag{12}$$

$$\sum p_i f_i \omega(p, \omega) = 1$$

From (11) and (9),

$$\begin{aligned} \forall j = 1, 2, \ldots, n \quad & v_j(p, \omega) \\ & = v\omega(p, \omega) \sum p_i f_{ij}(p, \omega) \end{aligned}$$

and from this and (12),

$$\begin{aligned} \forall j = 1, 2, \ldots, n \qquad & v_j(p, \omega) \\ & = -f_j(p, \omega) v\omega(p, \omega) \end{aligned} \tag{13}$$

Equation (13) is the main result connecting v with f. From (3), (a), (10) and (13) there follow

$$\forall j = 1, 2, \ldots, n \quad v_j(p, \omega) < 0 \tag{14}$$

and Roy's Identity (1942, p. 24; 1947, p. 217),

$$\begin{aligned} \forall j = 1, 2, \ldots, n \quad & x_j^* \\ & = -v_j(p, \omega) | v\omega(p, \omega) \end{aligned} \tag{15}$$

As deservedly famous as is (15), its equivalent version (13) reveals the structures involved more clearly, since it focuses sharply on the relations

between the functions v and f rather than the particular quantities x_j^*. In each of Roy's contributions the identity is first given in the form $v_j(p,\omega)|x_j^* = -v\omega(p,\omega)$, and is used primarily to prove (14); later, in Roy (1947, p. 220), the identity takes the more usual form (15).

Since (13) is an identity, differentiating it wrt any p_i yields $\forall i, j = 1, 2, \ldots, n$

$$-v_{ji}(p,\omega) = f_{ji}(p,\omega)v\omega(p,\omega) + f_j(p,\omega)v\omega_i(p,\omega)$$

Applying Young's Theorem to these equations, by symmetry,

$$\begin{aligned}\forall i,j = 1,2,\ldots,n\ & f_{ij}(p,\omega)v\omega(p,\omega) \\ & + f_i(p,\omega)v_{\omega j}(p,\omega) \\ & = f_{ij}(p,\omega)v_{\omega i}(p,\omega)\end{aligned} \tag{16}$$

Now make the quite restrictive assumption that for each p_i, $v_{\omega i}(p,\omega) = 0$; this requires in effect that each good have unitary elasticity of demand (see Samuelson, 1942, pp. 80–81). Then from (10) and (16).

$$\forall i,j = 1,2,\ldots,n\quad f_{ij}(p,\omega) = f_{ij}(p,\omega)$$

which are Slutsky-like equations that apply not to compensated but to ordinary demand functions.

Relations with the Cost Function and the Compensated Demand Functions

Suppose now that a *target* level of utility is specified and the following new optimization problem posed:

$$\begin{aligned}&\text{Min}(p,\ \tau) : \text{Find} \\ &x \in R^{n+}\ \text{to min}\ \langle p,x\rangle\ \text{subject to}\ u(x)\ \geq\ \tau\end{aligned} \tag{17}$$

Assume that a unique solution x^{**} to this problem exists, yielding a value $\langle p, x^{**}\rangle$. This implies the existence of two functions of the price-target pair (p, τ), called the *compensated* (or Hicksian) demand function h: $R^{n++} \times R \rightarrow R^{n+}$, defined by

$$x^{**} = h(p,\tau) \tag{18}$$

and the *cost (or expenditure)* function γ: $R^{n++} \times R \rightarrow R^{+}$, given by

$$\langle p^1, x^{**}\rangle = \gamma(p,\tau) \tag{19}$$

Retain assumptions (a)–(c), replacing x^* by x^{**}. Define M: $R^{n+} \times R^{n++} \rightarrow R$ by putting

$$M(x^1,p^1) = \gamma(p^1, u(x^1)) - \langle p^1, x^1\rangle \tag{20}$$

where x^1 and p^1 are arbitrary, as before. It follows that $M(\cdot, p^1)$ is negative semi-definite. Putting $p^1 = p$, the actual prices, it follows that if x^{**} solves Max (p, ω) then

$$M(x^{**},p) = \gamma(p,u(x^{**})) - \langle p,x^{**}\rangle = 0 \tag{21}$$

so that x^{**} maximizes $M(\cdot, p)$. Then a development *exactly* like that of the last section leads to a simple but basic result on the interrelations betwen h and γ, namely:

$$\forall j = 1,2,\ldots,n\quad \gamma_j(p,\tau) = h_j(p,\tau) \tag{22}$$

where h_j is the compensated demand function for the jth good. From (22) and (a),

$$\forall j = 1,2,\ldots,n\quad \gamma_j(p,\tau) > 0 \tag{23}$$

From (18), (22) can be rewritten in the more customary version that has come to be called *Shephard's Lemma* (Shephard 1953), although it dates back at least to Hotelling (1932).

$$\forall j = 1,2,\ldots,n\quad x_j^{**} = \gamma_j(p,\tau) \tag{24}$$

Thus (22) (or the Lemma) plays a role in the analysis of this problem which is symmetrical to that played by (13) (or Roy's Identity) in the analysis of Max(p, ω).

However, there are two important structural asymmetries between the problems max(p, ω)

and min(p, τ). First, suppose that for some reason (such as incompleteness of preferences) the utility function u does not exist, so that v does not exist either. Clearly, since Max(p, ω) requires a scalar measure of utility it cannot be defined in this new situation. However, by replacing the target level τ of utility by a target bundle x^τ, one can still define a perfectly sensible minimum problem Min(p, x^τ).

The second asymmetry is that while $v(\cdot, \omega)$ is only quasiconvex, $\gamma(\cdot, \tau)$is actually concave, and this without any assumptions on preferences. Since (full) concavity imposes sharper restrictions on any function than does quasi-convexity, the analysis of Min (p, τ) (or of Min (p, x^τ)) yields easier proofs of basic results than does that of Max (p, ω). For example, from (22) $h_{jj}(p, \tau) = \gamma_{jj}(p, \tau)$, and since $\gamma(\cdot, \tau)$ is concave $\gamma_{jj}(p, \tau) \leq 0$, proving that the substitution effect is non-positive.

Duality

It is not productive to oppose the virtues of minimum problems to those of maximum problems. Indeed, the most efficient path of the derivation of such propositions as the 'Fundamental Equation or Value Theory' (Hicks 1939, p. 309) is by a judicious mixture of the two, i.e. by first solving max(p, ω) to obtain $\tau^* = v(p, \omega)$ and $x^* = f(p, \omega)$, and then showing that x^* also solves min(p, τ^*). One interesting result that one can reach by this route relates all four functions v, f, γ and h in one equation:

$$\forall\ i, j = 1, 2, \ldots, n \qquad \gamma_i(p, \tau^*) f_j \omega(p, \omega) = -\ v_i(p, \omega) h_{j\tau}(p, \tau^*) \tag{25}$$

Since from Shephard's Lemma the left-hand side of (25) is the Hicksian income effect of a change in p_i on the demand for good j, so is the right-hand side (RHS). Notice that although each of the components of the RHS is affected by choice of the utility index u, their product is not.

Revert now to the assumptions of section "Definition and Simple Properties". The problems Max(p, ω) and Min (p, τ^*) are often referred to in the literature as dual to each other. For reasons given in detail in the entry on cost minimization and utility maximization, this usage seems inappropriate. However, as pointed out by Konyus and Byushgens (1926, p. 159) and Houthakker (1951–2, pp. 157–8), there *is* an interesting duality between the functions u and v. To show this, first rewrite the given prices and income (p, ω) as (p^*, ω^*), where $\omega^* > 0$ will be kept constant throughout. Next, define new income-normalized prices $q \in R^{n++}$ for any p by

$$q = (\omega^*) - 1p \tag{26}$$

Then use the homogeneity of f and v in (p, ω) to put them in the normalized forms F: $R^{n++} \to R^{n+}$ and w: $R^{n++} \to R$, defined by

$$F(q^*) \equiv f(p^*, \omega^*)$$

and

$$w(q^*) \equiv v(p^*, \omega^*) \tag{27}$$

Let

$$A(q^*) = \{x \in R^{n+} : \langle q^*, x\rangle \leq 1\} = A(p, \omega).$$

Then Max(p^*, ω^*) can also be written in a new form:

$$\text{Max}(q^*) : \text{Find}\ \ x \in A(q^*)\ \ \text{to}\ \ \max\ \ u(x).$$

The data of Max(q^*) are q^* and u. In the same way, the chosen bundle x^* and w are the data for a problem dual to Max(q^*). Let $B(x^*) = \{q \in R^{n+}; \langle x^*, q\rangle \leq 1\}$.Then the dual problem, situated in the space of normalized prices q, is

$$\text{Min}(x^*) : \text{Find}\ \ q \in B(x^*)\ \ \text{to}\ \ \min\ \ w(q).$$

A unique solution q^{**} to Min(x^*) (for which the strict quasi-convexity of w would suffice) implies the existence of two functions $\varphi : R^{n++} \to R^{n++}$ and $U : \varphi : R^{n++} \to R$, defined analogously to (3) and to (2)-cum-(4) by

$$q^{**} = \varphi(x^*) \tag{28}$$

$$U(x^*) = w(q^{**}). \tag{29}$$

By the construction of Min(x^*), $x^* \in A(q)$ for any q. So $w(q)$ must be at least as large as the utility level at x^*. But since x^* is bought at q^*, that utility level is $w(q^*)$.

Thus

$$\forall q \in B(x^*)\ w(q) \geq w(q^*) \tag{30}$$

Since min(x^*) is assumed to have a unique solution, (30) says that it must be q^*. If follows from this, (26) and (28) that φ is actually the inverse demand function F^{-1}. Moreover, $U(x^*) = w(q^*)$. So from this, (27), (4) and (2),

$$U(x^*) = u(x^*) \tag{31}$$

However, it cannot be concluded from (31) that $U \equiv u$ unless every bundle x in the domain of u is bought at some price-income pair (p, ω) and so can be an optimizing bundle such as x^*. This property requires that u be strictly quasiconcave. Granted that, (31) shows that the direct utility function u is recoverable from the indirect utility function w, just as w is obtainable from u.

See Also

- ▶ Demand Theory
- ▶ Index Numbers
- ▶ Roy, René François Joseph (1894–1977)

Bibliography

Antonelli, G.B. 1886. *Sulla teoria matematica della economia politica*. Pisa: Tipografia del Folchetto.

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

Court, L.M. 1941. Entrepreneurial and consumer demand theories for commodity spectra. *Econometrica* 9 (135–62): 241–297.

Diewert, W.E. 1982. Duality approaches to microeconomic theory. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. II. Amsterdam: North-Holland. Ch. 12, 535–599.

Fenchel, W. 1953. *Convex cones, sets and functions*. New Jersey: Department of Mathematics, Princeton University, mimeo.

Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of demand and supply functions. *Journal of Political Economy* 40: 577–616.

Houthakker, H.S. 1951. Compensated changes in quantities and qualities consumed. *Review of Economic Studies* 19: 155–164.

Houthakker, H.S. 1960. Additive preferences. *Econometrica* 28: 244–257.

Konyus, A.A. 1924. The problem of the true index of the cost of living. *Economic Bulletin of the Institute of Economic Conjuncture* 9–10: 64–71. Trans. in *Econometrica* 7, 1939, 10–29.

Konyus, A.A., and S.S. Byushgens. 1926. K probleme pokupatelnoi cili deneg. *Voprosi Konyunkluri* 2 (1): 151–172.

Malmquist, S. 1953. Index numbers and indifference surfaces. *Trabajos de estadistica* 4: 209–241.

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: 205–225.

Samuelson, P.A. 1942. Constancy of the marginal utility of income. In *Studies in mathematical economics and econometrics: In memory of Henry Schultz*, ed. O. Lange, F. McIntyre, and T.O. Yntema, 75–91. Chicago: University of Chicago Press.

Shephard, R.W. 1953. *Cost and production functions*. Princeton: Princeton University Press.

Varian, H.R. 1984. *Microeconomic analysis*. 2nd ed. New York: Norton.

Ville, J. 1946. Sur les conditions d'existence d'une ophelimité totale et d'un indice du niveau des prix. *Annales de l'Université de Lyon* 9: 32–39. Trans. in *Review of Economic Studies* 19, 1951, 123–128.

Wold, H.O.A. 1943–4. A synthesis of pure demand analysis. *Skandinavisk Aktuarietidskrift* 26: 85–144, 220–75; 27: 69–120.

Individual Learning in Games

Teck H. Ho

Abstract

This article reviews individual models of learning in games. We show that the experience-weighted attraction (EWA) learning nests different forms of reinforcement and belief learning, and that belief learning is mathematically equivalent to generalized reinforcement, where

even unchosen strategies are reinforced. Many studies consisting of thousands of observations suggest that the EWA model predicts behaviour out-of-sample better than its special cases. We also describe a generalization of EWA learning to investigate anticipation by some players that others are learning. This generalized framework links equilibrium and learning models, and improves predictive performance when players are experienced and sophisticated.

Keywords

Belief learning; Curse of knowledge; Equilibrium; Experience-weighted attraction (EWA) learning; Extensive-form games; Fictitious play; Forgone payoffs; Individual learning in games; Individual models of learning; Maximum likelihood; Mixed-strategy equilibrium; Noise; Overconfidence; Population models of learning; Quantal response equilibrium; Reinforcement learning; Signalling; Social calibration; Sophisticated players

JEL Classifications

C9

Introduction

Economic experiments on strategic games typically generate data that, in early rounds, violate standard equilibrium predictions. However, subjects normally change their behaviour over time in response to experience. The study of learning in games is about how this behavioural change works empirically. This empirical investigation also has a theoretical payoff: if subjects' behaviour converges to an equilibrium, the underlying learning model becomes a theory of equilibration. In games with multiple equilibria, this same model can also serve as a theory of equilibrium selection, a long-standing challenge for theorists.

There are two general approaches to studying learning: population models and individual models.

Population models make predictions about how the aggregate behaviour in a population will change as a result of aggregate experience. For example, in replicator dynamics, a population's propensity to play a certain strategy will depend on its 'fitness' (payoff) relative to the mixture of strategies played previously (Friedman 1991; Weibull 1995). Models like this submerge differences in individual learning paths.

Individual learning models allow each person to choose differently, depending on the experiences each person has. For example, in Cournot dynamics, subjects form a belief that other players will always repeat their most recent choice and best-respond accordingly. Since players are matched with different opponents, their best responses vary across the population. Aggregate behaviour in the population can be obtained by summing individual paths of learning.

This article reviews three major approaches to individual learning in games: experience-weighted attraction (EWA) learning, reinforcement learning, and belief learning (including Cournot and fictitious play). These models of learning strive to explain, for every choice in an experiment, how that choice arose from players' previous behaviour and experience. These models assume strategies have numerical evaluations, which are called 'attractions'. Learning rules are defined by how attractions are updated in response to experience. Attractions are then mapped into predicted choice probabilities for strategies using some well-known statistical rule (such as logit).

The three major approaches to learning assume players that are adaptive (that is, they respond only to their own previous experience and ignore others' payoff information) and that their behaviour is not sensitive to the way in which players are matched. Empirical evidence suggests otherwise. There are subjects who can anticipate how others learn and choose actions to influence others' path of learning in order to benefit themselves. So we describe a generalization of these adaptive learning models to allow for this kind of sophisticated behaviour. This generalized model assumes that there is a mixture of adaptive learners and sophisticated players. An adaptive learner adjusts his behaviour according to one of

the above learning rules. A sophisticated player does not learn and rationally best-responds to his forecast of others' learning behaviour. This model therefore allows 'one-stop shopping' for investigating the various statistical comparisons of learning and equilibrium models.

EWA Learning

Denote player i's jth strategy by s_i^j and the other player(s)' strategy by s_{-i}^k. The strategy actually chosen in period t is $s_i(t)$. Player i's payoff for choosing s_i^j in period t is $\pi_i\left(s_i^j, s_{-1}^k(t)\right)$. Each strategy has a numerical evaluation at time t, called an attraction $A_i^j(t)$. The model also has an experience weight, $N(t)$. The variables $N(t)$ and $A_i^j(t)$ begin with prior values and are updated each period. The rule for updating attraction sets $A_i^j(t)$ to be the sum of a depreciated, experience-weighted previous attraction $A_i^j(t-1)$ plus the (weighted) payoff from period t, normalized by the updated experience weight:

$$A_i^j(t) = \frac{\varphi \cdot N(t-1) \cdot A_i^j(t-1) + \left[\delta + (1-\delta) \cdot I\left(s_i^j, s_i(t)\right)\right] \cdot \pi_i\left(s_i^j, s_{-i}(t)\right)}{N(t)} \tag{2.1}$$

where indicator variable $I(x, y)$ is 1 if $x = y$ and 0 otherwise. The experience weight is updated by:

$$N(t) = \rho \cdot N(t-1) + 1. \tag{2.2}$$

Let $\kappa = \frac{\varphi - \rho}{\varphi}$. Then $\rho = \varphi \cdot (1 - \kappa)$ and $N(t)$ approaches the steady-state value of $\frac{1}{1-\varphi \cdot (1-\kappa)}$. If $N(0)$ begins below this value, it steadily rises, capturing an increase in the weight placed on previous attractions and a (relative) decrease in the impact of recent observations, so that learning slows down.

Attractions are mapped into choice probabilities using a logit rule (other functional forms fit about equally well; Camerer and Ho 1999):

$$P_i^j(t+1) = \frac{e^{\lambda \cdot A_i^j(t)}}{\sum_k e^{\lambda \cdot A_i^k(t)}}, \tag{2.3}$$

where λ is the payoff sensitivity parameter. The key parameters are δ, φ and κ (which are generally assumed to be in the [0,1] interval).

The most important parameter, δ, is the weight on forgone payoffs relative to realized payoffs. It can be interpreted as a kind of 'imagination' of forgone payoffs, or responsiveness to forgone payoffs (when δ is larger players move more strongly toward *ex post* best responses). We call it 'consideration' of forgone payoffs. The weight on forgone payoff δ is also an intuitive way to formalize the 'learning direction' theory of Selten and Stoecker (1986). Their theory consists of an appealing property of learning: subjects move in the direction of *ex post* best-response. Broad applicability of the theory has been hindered by defining 'direction' only in terms of numerical properties of ordered strategies (for example, choosing 'higher prices' if the *ex post* best response is a higher price than the chosen price). The parameter δ defines the 'direction' of learning set-theoretically by shifting probability towards the set of strategies with higher payoffs than the chosen ones.

The parameter φ is naturally interpreted as depreciation of past attractions, $A_i^j(t-1)$. In a game-theoretic context, φ will be affected by the degree to which players realize other players are adapting, so that old observations on what others did become less and less useful. So we can interpret φ as an index of (perceived) 'change' in the environment.

The parameter κ determines the growth rate of attractions, which in turn affects how sharply players converge. When $\kappa = 0$, the attractions are weighted averages of lagged attractions and payoff reinforcements (with weights $\varphi \cdot N(t-1)/(\varphi \cdot N(t-1) + 1)$ and $1/(\varphi \cdot N(t-1) + 1)$. When $\kappa = 1$ and $N(t) = 1$, the attractions are

cumulations of previous reinforcements rather than averages (that is, $A_i^j(t) = \varphi \cdot A_i^j(t-1) + \left[\delta + (1-\delta) \cdot I\left(s_i^j, s_i(t)\right)\right] \cdot \pi_i\left(s_i^j, s_{-i}(t)\right)\Big)$. In the logit model, the *differences* in strategy attractions determine their choice probabilities. When κ is high the attractions can grow furthest apart over time, making choice probabilities closer to zero and one. We therefore interpret κ as an index of 'commitment'.

Reinforcement Learning

In cumulative reinforcement learning (Harley 1981; Roth and Erev 1995), strategies have levels of attraction which are incremented by only received payoffs. The initial reinforcement level of strategy j of player i, s_i^j, is $R_i^j(0)$. Reinforcements are updated as follows:

$$R_i^j(t) = \begin{cases} \varphi \cdot R_i^j(t-1) + \pi_i\left(s_i^j, s_{-i}(t)\right) & \text{if } s_i^j = s_i(t), \\ \varphi \cdot R_i^j(t-1) & \text{if } s_i^j \neq s_i(t). \end{cases} \tag{3.1}$$

Using the indicator function, the two equations can be reduced to one:

$$R_i^j(t) = \varphi \cdot R_i^j(t-1) + I\left(s_i^j, s_i(t)\right) \cdot \pi_i\left(s_i^j, s_{-i}(t)\right). \tag{3.2}$$

This updating formula is a special case of the EWA rule, when $\delta = 0$, $N(0) = 1$, and $\kappa = 1$.

In average reinforcement learning, updated attractions are *averages* of previous attractions and received payoffs (for example, Mookerjhee and Sopher 1994; 1997; Erev and Roth 1998). For example

$$R_i^j(t) = \varphi \cdot R_i^j(t-1) + (1-\varphi) \cdot I\left(s_i^j, s_i(t)\right) \cdot \pi_i\left(s_i^j, s_{-i}(t)\right). \tag{3.3}$$

A little algebra shows that this updating formula is also a special case of the EWA rule, when $\delta = 0$, $N(0) = \frac{1}{1-\varphi}$, and $\kappa = 0$. Since the two reinforcement models are special cases of EWA learning, their predictive adequacy can be tested empirically by setting the appropriate EWA parameters to their restricted values and seeing how much fit is compromised (adjusting, of course, for degrees of freedom).

Belief Learning

In belief-based models, adaptive players base their responses on beliefs formed by observing their opponents' past plays. While there are many ways of forming beliefs, we consider a fairly general 'weighted fictitious play' model, which includes fictitious play (Brown 1951; Fudenberg and Levine 1998) and Cournot best-response (Cournot 1960) as special cases. It corresponds to Bayesian learning if players have a Dirichlet prior belief.

In weighted fictitious play, players begin with prior beliefs about what the other players will do, which are expressed as ratios of strategy choice counts to the total experience. Denote total experience by $N(t) = \sum_k N_{-i}^k(t)$. Express the belief that others will play strategy k as $B_{-i}^k(t) = \frac{N_{-i}^k(t)}{N(t)}$, with $N_{-i}^k(t) \geq 0$ and $N(t) > 0$.

Beliefs are updated by depreciating the previous counts by φ, and adding one for the strategy combination actually chosen by the other players. That is,

$$B_{-i}^k(t) = \frac{\varphi \cdot N_{-i}^k(t-1) + I\left(s_{-i}^k, s_{-i}(t)\right)}{\sum_h \left[\varphi \cdot N_{-i}^h(t-1) + I\left(s_{-i}^h, s_{-i}(t)\right)\right]}. \tag{4.1}$$

This form of belief updating weights the belief from one period ago φ times as much as the most recent observation, so φ can be interpreted as how quickly previous experience is discarded. When $\varphi = 0$ players weight only the most recent observation (Cournot dynamics); when $\varphi = 1$ all previous observations count equally (fictitious play).

Given these beliefs, we can compute expected payoffs in each period t,

$$E_i^j(t) = \sum_k B_{-i}^k(t)\pi\left(s_i^j, s_{-i}^k\right). \quad (4.2)$$

The crucial step is to express period t expected payoffs as a function of period $t - 1$ expected payoffs. This yields:

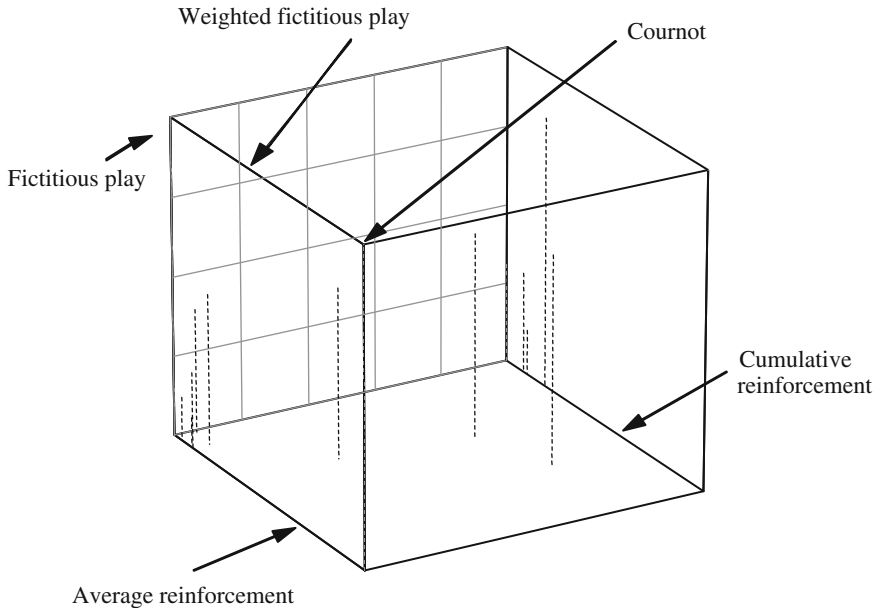
$$E_i^j(t) = \frac{\varphi \cdot N(t-1) \cdot E_i^j(t-1) + \pi\left(s_i^j, s_{-i}(t)\right)}{\varphi \cdot N(t-1) + 1}. \quad (4.3)$$

By expressing expected payoffs as a function of lagged expected payoffs, we make the belief terms disappear. This is because the beliefs are only used to compute expected payoffs, and when beliefs are formed according to weighted fictitious play, the expected payoffs which result can also be generated by generalized reinforcement according to previous payoffs. More precisely, if the initial attractions in the EWA model are expected payoffs given some initial beliefs (that is, $A_i^j(0) = E_i^j(0)$, $\kappa = 0$ (or $\varphi = \rho$), and foregone payoffs are weighted as strongly as received payoffs ($\delta = 1$), then EWA attractions are *exactly* the same as expected payoffs. Put differently, belief learning is 'mathematically equivalent' or 'observationally equivalent' to EWA learning with $\delta = 1$, $\kappa = 0$ and $A_i^j(0) = E_i^j(0)$.

This demonstrates a close kinship between reinforcement and belief approaches. Belief learning is nothing more than generalized attraction learning in which strategies are reinforced equally strongly by actual payoffs and foregone payoffs and attractions are weighted averages of past attractions and reinforcements. Hopkins (2002) compares the convergence properties of reinforcement and fictitious play and finds that they are quite similar in nature and that they will in many cases have the same asymptotic behaviour.

A Graphical Representation

Since reinforcement and belief learning are special cases of EWA learning, it is possible to represent all three learning models in a three-dimensional EWA cube (see Fig. 1). The vertex $\delta = 1$ and $\kappa = 0$ corresponds, to weighted fictitious play models. The corners $\varphi = 0$ and $\varphi = 1$



Individual Learning in Games, Fig. 1 EWA's model parametric space

correspond to Cournot best-response dynamics and fictitious play, respectively. Reinforcement models in which only chosen strategies are reinforced according to their payoffs correspond to vertices in which $\varphi = 0$, and $\kappa = 1$ (cumulative reinforcement) or $\kappa = 0$ (averaged reinforcement). Interior configurations of parameter values incorporate both the intuition behind reinforcement learning, that realized payoffs weigh most heavily ($\delta < 1$), and the intuition implicit in belief learning, that foregone payoffs matter too ($\delta > 0$).

The cube shows that contrary to popular belief for many decades, reinforcement and belief learning are simply two extreme configurations on opposite edges of a three-dimensional cube, rather than fundamentally unrelated models. Figure 1 also shows estimates of the three parameters in 20 different studies (Camerer et al. 2002). Each point is a triple of estimates. These parameter estimates were typically obtained by the maximum likelihood method. Initial attractions could be either estimated using data or set to plausible values using the cognitive hierarchy model of one-shot games; see Camerer et al. (2004) for details. Most points are sprinkled throughout the cube, rather than at the extreme vertices mentioned in the previous paragraph, although some (generally from games with mixed-strategy equilibria) are near the averaged reinforcement corner $\delta = 0$ and $\kappa = \varphi = 1$. Ho et al. (2007) provide an explanation for how δ and φ vary across games by endogenizing them as functions of game experience. Parameter estimates are generally significantly inside the interior of the cube rather than near the vertices. Thus, we may conclude that subjects' behaviour is often neither belief nor reinforcement learning.

Linking Learning and Equilibrium Models

The adaptive learning models presented above do not permit players to anticipate learning by others. Omitting anticipation logically implies that players do not use information about the payoffs of other players, and that whether players are matched together repeatedly or are randomly re-matched should not matter. Both of the latter implications are unintuitive, and experiments with experienced subjects have provided evidence to show otherwise.

In Camerer et al. (2002) and Chong et al. (2006), we proposed a simple way to include 'sophisticated' anticipation by some players that others are learning, using two additional parameters. We assume a fraction α of players are sophisticated. Sophisticated players think that a fraction $(1 - \alpha')$ of players are adaptive and the remaining fraction α' of players are sophisticated like themselves. They use the EWA model (which nests reinforcement and belief learning as special cases) to forecast what the adaptive players will do, and choose strategies with high expected payoffs given their forecast.

All the adaptive models discussed above (EWA, reinforcement, belief learning) are special cases of this generalized model with $\alpha = 0$. The assumption that sophisticated players think some others are sophisticated creates a small whirlpool of recursive thinking which implies that quantal response equilibrium (QRE; McKelvey and Palfrey 1995) and Nash equilibrium are special cases of this generalized model. Our specification also shows that equilibrium concepts combine two features which are empirically and psychologically separable: 'social calibration' (accurate guesses about the fraction of players who are sophisticated, $\alpha = \alpha'$); and full sophistication ($\alpha = 1$). Psychologists have identified systematic departures from social calibration called 'false uniqueness' or overconfidence ($\alpha > \alpha'$) and 'false consensus' or curse of knowledge ($\alpha > \alpha'$).

Formally, adaptive learners follow the EWA updating equations given above (that is, (2.1) and (2.2)). Sophisticated players have attractions $B_i^j(t)$ and choice probabilities $Q_i^j(t+1)$ specified as follows:

$$B_i^j(t) = \sum_k \left[(1-\alpha') \cdot P_{-i}^k(t+1) + \alpha' Q_{-i}^k(t+1)\right] \cdot \pi_i\left(s_i^j, s_{-i}^k\right), \tag{6.1}$$

$$Q_i^j(t+1) = \frac{e^{\lambda \cdot B_i^j(t)}}{\sum_k e^{\lambda \cdot B_i^k(t)}}. \tag{6.2}$$

The generalized model has been applied to experimental data from ten-period p-beauty contest games (specific details of data collection are given in Ho et al. 1998). In these games, seven subjects choose numbers in [0,100] simultaneously. The subject whose number is closest to p times the average (where $p = .7$ or .9) wins a fixed prize. Subjects playing for the first time are called 'inexperienced'; those playing another ten-period game (with a different p) are called 'experienced'.

The estimation results show that for inexperienced subjects, adding sophistication to adaptive EWA improves log likelihood (LL) substantially both in-and out-of-sample. The estimated fraction of sophisticated players is $\hat{\alpha} = .24$ and their estimated perception $\hat{\alpha}' = 0$. Experienced subjects show a much larger improved fit from sophistication, and a larger estimated proportion, $\hat{\alpha} = .75$. Their perceptions are again too low, $\hat{\alpha}' = .41$, showing a degree of overconfidence. The increase in sophistication due to experience reflects a kind of 'learning about learning', which is similar to rule learning (that is, subjects switch their learning rule over time (Stahl 2000; Ho et al. 2007). Overall, these results suggest that subjects are not socially calibrated, that not all subjects are sophisticated, and that the proportion of sophistication grows with experience.

Conclusions and Future Research

We describe three major approaches of adaptive learning models. We show that EWA learning is a generalization of reinforcement and belief learning and that the latter two nested models are intimately related. Specifically, they differ mainly in the way they treat forgone payoffs; reinforcement learning ignores them and belief learning treats them the same as actual payoffs. Estimation results from dozens of studies show that the emergence of behaviour is neither reinforcement nor belief learning in most games. The EWA cube provides a simple way for detecting how these simpler models fail and why.

We also describe a generalization of these adaptive models to study anticipation by some players that others are learning. This generalized model nests equilibrium and the adaptive learning models as special cases and is a powerful framework for analysing both equilibrium and learning simultaneously. We show that it can improve the predictive performance of the adaptive learning models when players are experienced and able to anticipate how others learn.

There are three promising areas of future research, all of which aim to make the above learning models more amenable to field applications.

1. *Transfer of learning across similar games.* In practice, it is unreasonable to expect people play the identical game again and again. Since people are more likely to face with similar but non-identical strategic situations, it is important to determine whether they are able to transfer learning from one situation to another. Cooper and Kagel (2004) provide evidence that subjects who have learned to play strategically in one signalling game can transfer most of this knowledge to related games. This transfer of learning occurs because the proportion of sophisticated players grows with experience (just like what we observed in p-beauty contest games discussed above). This positive evidence is encouraging but more work is necessary to determine whether this finding indeed generalizes to other games.
2. *Learning in extensive-form games.* Most of the learning literature focuses on strategic or normal-form games (for an exception see Anderson and Camerer 2000). This is done in part to simplify the learning context to situations where each action unambiguously corresponds to a final outcome. In extensive-form games or many field settings, where a final outcome is typically a result of a series of actions taken sequentially over time, there is a natural question how an action step taken at a particular time contributes to the final outcome. This 'credit assignment' problem is important because different agents might be responsible for different action steps, and some steps might be more crucial than others at determining the final outcome. A good

learning model should assign credit appropriately to each action step.

3. *Learning in noisy experiments.* There is a general belief that, given a sufficiently high stake and that people play repeatedly with a clear feedback, their behaviour will converge to equilibrium in the long run. However many real-world environments provide noisy feedback. So it is important to study how noise in feedback affects rates of learning and the likelihood of convergence to equilibrium.

See Also

- ▶ Experimental Economics
- ▶ Learning and Evolution in Games: Belief Learning
- ▶ Maximum Likelihood

Bibliography

Anderson, C., and C. Camerer. 2000. Experience-weighted attraction learning in sender-receiver signaling games. *Economic Theory* 16: 689–718.

Brown, G. 1951. Iterative solution of games by fictitious play. In *Activity analysis of production and allocation*. New York: Wiley.

Camerer, C.F., and T.-H. Ho. 1999. Experience-weighted attraction learning in normal-form games. *Econometrica* 67: 827–874.

Camerer, C., T.-H. Ho, and J.-K. Chong. 2002. Sophisticated learning and strategic teaching. *Journal of Economic Theory* 104: 137–118.

Camerer, C.F., T.-H. Ho, and J.-K. Chong. 2004. A cognitive hierarchy model of one-shot games. *Quarterly Journal of Economics* 119: 861–898.

Chong, J.-K., C. Camerer, and T.-H. Ho. 2006. A learning-based model of repeated games with incomplete information. *Games and Economic Behavior* 55: 340–371.

Cooper, D., and Kagel, J. 2004. Learning and transfer in signaling games. Working paper. Ohio State University.

Cournot, A. 1960. *Recherches sur les principes mathématiques de la théorie des richesses. Trans. N. Bacon as researches in the mathematical principles of the theory of wealth*. Haffner: London.

Erev, I., and A. Roth. 1998. Modelling predicting how people play games: Reinforcement learning in experimental games with unique, mixed-strategy equilibria. *American Economic Review* 88: 848–881.

Friedman, D. 1991. Evolutionary games in economics. *Econometrica* 59: 637–666.

Fudenberg, D., and D. Levine. 1998. *The theory of learning in games*. Cambridge, MA: MIT Press.

Harley, C. 1981. Learning the evolutionary stable strategies. *Journal of Theoretical Biology* 89: 611–633.

Ho, T.-H., C. Camerer, and K. Weigelt. 1998. Iterated dominance and iterated best-response in *p*-beauty contests. *American Economic Review* 88: 947–969.

Ho, T.-H., C. Camerer, and J.-K. Chong. 2007. Self-tuning experience-weighted attraction learning in games. *Journal of Economic Theory* 133: 177–198.

Hopkins, E. 2002. Two competing models of how people learn in games. *Econometrica* 70: 2141–2166.

McKelvey, R., and T. Palfrey. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10: 6–38.

Mookerjhee, D., and B. Sopher. 1994. Learning behavior in an experimental matching pennies game. *Games and Economic Behavior* 7: 62–91.

Mookerjee, D., and B. Sopher. 1997. Learning and decision costs in experimental constant-sum games. *Games and Economic Behavior* 19: 97–132.

Roth, A.E., and I. Erev. 1995. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* 8: 164–212.

Selten, R., and R. Stoecker. 1986. End behavior in sequences of finite Prisoner's Dilemma supergames: A learning theory approach. *Journal of Economic Behavior and Organization* 7: 47–70.

Stahl, D. 2000. Rule learning in symmetric normal-form games: Theory and evidence. *Games and Economic Behavior* 32: 105–138.

Weibull, J. 1995. *Evolutionary game theory*. Cambridge, MA: MIT Press.

Individual Retirement Accounts

Jonathan Skinner

Abstract

Individual Retirement Accounts in the United States are tax-preferred saving vehicles designed to encourage saving for retirement. Many countries have adopted similar saving mechanisms such as Individual Saving Accounts in Britain, Special Saving Incentive Accounts in Ireland, and Tax-Preferred Deposit Accounts in Belgium. Enrolment rates are substantially higher among high income taxpayers, while the saving effects are

often found to be quite modest. However, given the often low revenue cost of these tax preferred accounts, they may be reasonably cost-effective in terms of new saving per lost unit of revenue.

Keywords

Capital gains; Individual Retirement Accounts; National saving; Precautionary saving; Retirement; Pensions

JEL Classifications

H2

The Individual Retirement Account (IRA) in the United States was first introduced in 1974, but languished in relative obscurity until the Economic Recovery Tax Act of 1981 expanded eligibility to all US taxpayers. Contributions jumped from $4.8 billion in 1981 to $28.3 billion in 1982, before peaking at $37.8 billion in 1986 (Holden et al. 2005). The traditional IRA provided a tax break when contributions were made to qualified accounts, but taxed the entire withdrawal (principle plus interest) upon withdrawal. Restrictions included a ten per cent penalty for withdrawing money before age $59\frac{1}{2}$, and the requirement that the taxpayer implement a systematic withdrawal plan by age $70\frac{1}{2}$. In 2007, the limit for tax-deductible contributions was $4,000, or $5,000 for taxpayers over age 50.

The IRA came under fire during the mid-1980s because of revenue costs and concerns that it was being used as a tax shelter for high-income taxpayers. The Tax Reform Act of 1986 tax instituted income limits and, as a result, contributions dropped off rapidly, from 37.8 billion in 1986 to 14.1 billion in 1987 (Holden et al. 2005). That contributions even fell by 30 per cent among those still eligible to contribute suggests that confusion about eligibility (Hrung 2001), or a decline in advertising, may have affected taxpayer participation adversely. While the introduction of the eponymous Roth IRA, under which the taxpayer contributed after-tax dollars which were allowed to accumulate (and be withdrawn) tax-free, was popular among contributors, IRAs remain a relatively unimportant source of new saving, accounting for less than 0.2 per cent of GDP.

Nonetheless, the stock of IRA assets had grown to $3 trillion by 2007, when it comprised 20 per cent of US retirement saving (Holden et al. 2005). The reason IRAs comprise such a large fraction of wealth is that workers changing jobs or retiring are allowed to 'roll over' defined contribution (401(k)) balances into IRAs without any tax penalty. Thus IRA growth has been fuelled by these rollovers, which in 2000–1 comprised more than $200 billion annually, or about ten times the contributions by savers (Holden et al. 2005).

A number of IRA-like saving vehicles have been introduced in other countries, often with more generous eligibility and contribution rules. For example, the current (2007) contribution limit for Registered Retirement Saving Plans (RRSPs) in Canada is C$19,000. As well, taxpayers may carry forward past unused contributions, so the effective limit is generally much larger. In the United Kingdom, Individual Saving Accounts (ISAs) were introduced in 1999, replacing Personal Equity Plans (PEPs) and Tax-Exempt Special Saving Accounts (TESSAs) (Attanasio, Banks and Wakefield 2004). The contribution limit for the ISA in 2007 was £7,000 and resembled a Roth IRA in that contributions were made after taxes were paid but withdrawals and accumulated build-up were tax-free. Many other developed countries offer similar tax incentives, such as tax-preferred saving accounts for children and grandchildren in Denmark, Special Saving Incentive Accounts in Ireland, and Tax-Preferred Deposit Accounts in Belgium (see Maffini 2007). Other tax-preferred saving schemes, most notably employer-based defined-contribution pension plans such as 401(k)s in the United States, are discussed elsewhere (see pensions).

Economic Incentives

As noted above, there are two basic flavours of IRAs, traditional IRAs with an 'up-front' deduction and Roth-style IRAs, whereby taxpayers

invest tax-free dollars and withdraw the accumulated amount tax-free. The economic effects of IRAs are simplest to see in the case of a standard bond that pays a constant rate of return r^* for n years until retirement when the entire IRA is withdrawn. The marginal tax rate on an extra dollar of interest income is τ_m, at which point the tax rate shifts to τ_o while retired. If the investor invests one dollar in the conventional bond, her after-tax return at retirement will be $(1 + r^*(1-\tau_m))^n$, while an investment in a classic IRA will yield

$$\frac{(1-\tau_o)(1+r^*)^n}{1-\tau_m}$$

and a Roth IRA will return $(1 + r^*)^n$. It is straightforward to demonstrate that both the classic and the Roth IRA strictly dominate the conventional bond investment, and that the classic IRA dominates the Roth IRA under the assumption that $\tau_o < \tau_m$. This may not always be a sound assumption, particularly if retirees are too worried about higher future tax rates to pay for financially strained social insurance programmes. Gokhale et al. (2001) have observed that in some cases taxable income while retired may be subject to a higher marginal tax rate because of peculiarities in the US tax code, diminishing the advantage of traditional IRAs relative to Roth IRAs.

The decision becomes more complex when considering whether to hold equity investments inside an IRA. When a substantial fraction of the asset appreciation occurs through capital gains, one must trade off the tax advantages of the conventional IRA with the necessity to withdraw the (appreciated) assets from the IRA account starting in age $70\frac{1}{2}$ (Note that the Roth IRA does not require a withdrawal plan). As well, keeping the stock outside of an IRA retains its availability for precautionary purposes, and makes it eligible for preferential treatment of capital gains and dividends, and the possibility of stepping up the tax basis at death.

There are two key reasons why countries may decide to create IRA-style accounts. The first is to stimulate national saving, while the second is to improve the financial security of retirees, particularly those without access to employer-based pensions. The two are not necessarily overlapping. A programme that successfully stimulates saving among millionaires and billionaires may have a large impact on aggregate national saving, but do little or nothing to enhance the financial security of these households already well-prepared for the risks of retirement. Similarly, a programme that encourages low-or lower-middle-class savers by supplementing financial resources by (say) $10,000 would have a small impact on national saving but could exert a much larger proportional impact on available financial resources. That IRA inflows (excluding rollovers) in the United States comprise less than 0.2 per cent of GDP suggests a small upper limit for its impact on aggregate saving. (The size of these plans in other countries, relative to GDP, also appears modest; see Maffini 2007.)

Whether as a mechanism to increase aggregate saving or to encourage retirement security for specific households, the impact of IRAs on net saving is theoretically ambiguous. If IRA wealth and non-IRA taxable wealth were perfect substitutes, clever taxpayers could simply shuffle money from their taxable wealth accounts into IRAs, and enjoy the future or current tax rebate. If the tax incentive is further financed through deficit spending, and taxpayers spent part of the tax break, net national saving could *decline* following the introduction of an IRA programme. How individual accounts affect individual and national saving is therefore an empirical question.

Empirical Evidence

There has been considerable debate regarding the impact of IRAs on net saving. The first set of studies was by Venti and Wise (1986, 1990) who estimated that IRA and non-IRA savings were imperfect substitutes, thus suggesting that IRAs led to roughly 60 cents of new saving per dollar of IRA contributions, with most of the remaining 40 cents representing the tax subsidy. Similarly,

Engelhardt (1996) found large saving effects by comparing saving rates in Canada before and after the cessation of the tax-subsidized Registered Home Ownership Savings Plan.

Gale and Scholz (1994) specified a more general model allowing for differences in tastes for saving between IRA and non-IRA contributions, and using the nonlinearity of the budget constraint – due to the IRA limits – to help identify the true saving effects of IRAs. They arrived at a quite different conclusion, namely, that IRAs in fact reduced net saving because contributors 'shuffled' savings from taxable accounts. Ultimately, their estimates were found to be very sensitive to the inclusion or exclusion of a few observations (Poterba et al. 1996), underscoring the difficultly of testing for causality using observational data. Even in a dynamic setting (as in Feenberg and Skinner 1989) one cannot rule out the possibility that former spendthrifts who suddenly start pouring money into IRAs would have done so even without IRAs available for their use.

Attanasio et al. (2004) used the natural experiment of the 1999 shift in the United Kingdom from PEPs and TESSAs to the less restrictive (and hence more popular) ISAs to test the resulting impact on saving. The resulting (albeit very noisy) patterns of changes in saving rates were not supportive of a positive impact on national saving. Another study used the difference between contribution rates of taxpayers making their first year's contribution to an IRA and those of later contributors (Attanasio and De Leire 2002; see also Joines and Manegold 1995). They found that new contributors exhibit shuffling behaviour from existing assets into IRAs. Less clear is whether later contributors (the majority of IRA inflows) were increasing net national saving (Hubbard and Skinner 1996; Attanasio et al. 2004).

The strongest evidence of how IRAs affect saving comes from a randomized trial conducted in the St. Louis metropolitan area by H&R Block, a large tax preparation firm (Duflo et al. 2006). In this study, tax filers at H&R Block were provided with different incentives to open an 'express' IRA funded with either tax refunds or other sources. Duflo et al. found enrolment rates of 3 per cent for the control group, 10 per cent for the treatment group with a 20 per cent match and 17 per cent for those with the 50 per cent match. Conditional on enrolment, contributions (excluding the match) amounted to $860, $1,280, and $1,310, respectively. The researchers were not able to measure offsetting effects for non-IRA wealth, but large and significant effects in the treatment group were observed even in households with low median income or without saving accounts, thus minimizing the potential for 'shuffling' from other assets within these groups.

However, these results cannot be generalized to the saving effects of conventional IRAs in the United States or in other countries. As the authors noted, the treatment effect depended strongly on the specific tax professional; some tax professionals just couldn't 'sell' the IRAs no matter how attractive the match. Furthermore, the IRA was offered at an auspicious time when the refund had not yet been issued. The IRA match may have been a necessary, but apparently it was not a sufficient, condition to persuade all contributors (even those in high income brackets) to sign up.

Conclusions

It is unfortunate that we still know so little about the saving effects of IRAs and similar saving incentives. While we don't know the incremental effect of IRAs, we might expect the strongest saving effects to arise among lower-income households where the opportunity to shuffle assets is most constrained (for example, Engen and Gale 2000; Benjamin 2005 in the case of 401(ks)). And the evidence we do observe the sharp drops in contributions among those still eligible following the 1986 cutbacks, and the importance of individual tax professional effects, for example – suggests that behavioural or marketing factors are critically important in 'selling' IRAs to the households where the tax advantages are perhaps no so apparent and the distributional effects of the subsidy are not so inequitable (see Bernheim 1997).

What about the saving effects of IRAs among higher-income households? Because IRA accounts typically require the enrolee to write a check, one may never expect it to exhibit the same saving effects of a 401(k) plan that automatically withdraws money before the paycheck is cashed. But, as Hubbard and Skinner (1996) argue, the saving effects need not be large in order to justify the government provision of saving accounts. Recall that the net revenue loss to the government for the traditional IRA is the up-front deduction, less the present value of the discounted future tax payments. This difference may be quite modest when strong stock market gains build up equity inside traditional IRAs, leading to higher future revenue collections as the IRAs are gradually drawn down (Dusseault and Skinner 2000; also see Gravelle 2000).

In sum, IRAs provide tax-preferred wealth accumulation to those without employer pensions or who seek to accumulate something extra for retirement. Saving effects are likely to be largest when IRAs are designed to appeal to low-or middle-income households where opportunities for shuffling are minimized. Finally, governments may find policy changes irresistible as they realize how much future tax revenue lies within traditional IRA assets, or how much potential tax liability lies within rapidly growing Roth or ISA assets.

See Also

- ▶ Capital Gains Taxation
- ▶ Pensions
- ▶ Retirement
- ▶ Taxation of Income
- ▶ Taxation of Wealth

Bibliography

Attanasio, O., J. Banks, and M. Wakefield. 2004. Effectiveness of tax incentives to boost (retirement) saving: Theoretical motivation and empirical evidence. *OECD Economic Studies* 39(2): 145–172.

Attanasio, O., and T. De Leire. 2002. IRAs and household savings revisited: Some new evidence. *Economic Journal* 112: 504–538.

Benjamin, D. 2005. Does 401(k) eligibility increase saving? Evidence from propensity score subclassification. *Journal of Public Economics* 87: 1259–1290.

Bernheim, B.D. 1997. Rethinking savings incentives. In *Fiscal policy: Lessons from Economic Research*, ed. A. Auerbach. Cambridge: MIT Press.

Duflo, E., W.G. Gale, J.B. Liebman, P.R. Orszag, and E. Saez. 2006. Saving incentives for low-and middle-income families: Evidence from a field experiment with H&R block. *Quarterly Journal of Economics* 121: 1311–1346.

Dusseault, B., and J. Skinner. 2000. Did individual retirement accounts actually raise revenue? *Tax Notes* 86: 851–856.

Engelhardt, G. 1996. Tax subsidies and household saving: Evidence from Canada. *Quarterly Journal of Economics* 111: 1237–1268.

Engen, E.M., and W.G. Gale. 2000. The effects of 401(k) plans on households wealth: differences across earnings groups. Working Paper No. 8032. Cambridge, MA: NBER.

Engen, E.M., W.G. Gale, and J.K. Scholz. 1996. The illusory effects of saving incentives on saving. *Journal of Economic Perspectives* 10(4): 113–138.

Feenberg, D. and J. Skinner. 1989. Sources of IRA saving. In *Tax policy and the economy,* vol. 3, ed. L.H. Summer. Cambridge, MA: MIT Press.

Gale, W.G., and J.K. Scholz. 1994. IRAs and household saving. *American Economic Review* 84: 1233–1260.

Gokhale, J., L.J. Kotlikoff, and T. Neumann. 2001. Does participating in a 401(k) raise your lifetime taxes? Working Paper No. 8341. Cambridge, MA: NBER.

Gravelle, J. 2000. IRAs as revenue-raisers: A critique. *Tax Notes* 86(9).

Holden, S., K. Ireland, V. Leonard-Chambers, and M. Bogdan. 2005. The individual retirement account at age 30: A retrospective. *Investment Company Institute Perspective* 11(1): 1–23.

Hrung, W. 2001. Information and IRAs: The influence of tax preparers. *Journal of Public Economics* 80: 467–484.

Hubbard, G.R., and J.S. Skinner. 1996. Assessing the effectiveness of saving incentives. *Journal of Economic Perspectives* 10(4): 73–90.

Joines, D.H., and J.G. Manegold. 1995. *IRA and saving: Evidence from a panel of taxpayers*. Mimeo: University of Southern California.

Maffini, G. 2007. *Encouraging saving through tax-preferred accounts.* Tax policy studies no. 15. Paris: OECD.

Poterba, J., S. Venti, and D. Wise. 1996. How retirement saving programs increase saving. *Journal of Economic Perspectives* 10(4): 91–112.

Venti, S., and D. Wise. 1986. Tax-deferred accounts, constrained choice, and estimation of individual saving. *Review of Economic Studies* 53: 579–601.

Venti, S., and D. Wise. 1990. Have IRAs increased US saving? Evidence from consumer expenditure surveys. *Quarterly Journal of Economics* 105: 661–689.

Individualism

C. B. Macpherson

Individualism is social theory or ideology which assigns a higher moral value to the individual than to the community or society, and which consequently advocates leaving individuals free to act as they think most conducive to their self-interest. The term was also, as noted below, sometimes used in the 19th century as a name for an actual economic system. When so used, the term denoted the competitive market system which lets the direction of the economy be the unintended outcome of the decisions made by myriad individuals about the uses to which they will put their own labour and resources.

The first edition (1896) of Palgrave's *Dictionary of Political Economy* defined individualism in the latter, narrower sense. The article entitled Individualism began by reporting that John Stuart Mill had applied the term to 'that system of industrial organisation in which all initiative is due to private individuals, and all organisation to their voluntary agreement'. The article then remarked: 'The natural antithesis to individualism is COLLECTIVISM or we may say SOCIALISM, a system under which industry is directly organized by the state, which owns all means of production and manages all processes by appointed officers.' The author defined the fundamentals of the system of individualism quite precisely:

The essential features of individualism are, (1) private property in capital, to which are added almost of necessity the rights of bequest and inheritance, thus permitting unlimited transfer and accumulation. (2) competition, a rivalry between individuals in the acquisition of wealth, a struggle for existence in which the fittest survive.

There could hardly be a better definition of capitalism, at least of the neo-classical economists' model of capitalism. John Stuart Mill's *Socialism* is cited as authority for such a use of 'individualism', properly enough: his *Chapters on Socialism* (1879) does describe 'the principle of individualism' as 'competition, each one for himself and against all the rest. It is grounded on the opposition of interests, not the harmony of interests, and under it everyone is required to find his place by struggle, by pushing others back or being pushed back by them'; and later in the same work individualism is equated with 'quarrelling about material interests'. One might also cite Mill's earlier (1851) 'Newman's Political Economy', where 'the existing individualism', described as 'arming one human being against another, making the good of each depend upon evil to others', is said to be so morally inferior to socialism that socialism is 'easily triumphant' over it.

It may be thought that the Palgrave definition of individualism is unduly narrow: a modern scholar (Lukes 1973) has distinguished no less than eleven meanings the term may have, ranging from respect for human dignity, autonomy, privacy, and self-development, to epistemological and methodological individualism. Most of these meanings are indeed not considered in Palgrave's *Dictionary*, but since it is a dictionary *of political economy*, only meanings with an economic connotation can be expected to be treated. However, although that charge of undue narrowness may be dismissed, it may still appear that, considered historically, his usage is too narrow to be accurate for the whole modern Western tradition down to his own time.

The idea that the individual is morally more important than society goes back of course, in modern times, to the Renaissance. The same view, in religious terms, emerged at the Reformation, which made each individual, rather than the Church, the guardian of his own salvation; and this view got wider currency in 17th-century Puritanism. Neither the Renaissance nor the Reformation and the subsequent Puritanism reduced individuals to atoms of matter in motion, each seeking power and wealth at the expense of every other one. That step was taken by Hobbes in the mid-17th century: in his view, society was simply a congeries of colliding atoms in

unceasing motion. That puts Hobbes's individualism close to, but leaves it broader than, Palgrave's concept.

In the 18th century Adam Smith gave full market individualism a more pleasant face, arguing that the most beneficent possible social result would be attained by leaving individuals free to make self-interested bargains in a competitive market: that was the doctrine of *laissez-faire*. And the market economy was solidly enough established in England by Smith's time that it could be accepted as a part of the natural order by that venerator of the traditional hierarchy of ranks, Smith's contemporary, Edmund Burke, though in Burke's hands the market economy became a much less pleasant affair. His *Thoughts and Details on Scarcity* (1795) was an unqualified endorsement of *laissez-faire*: it issued a shrill warning against 'breaking the laws of commerce, which are the laws of nature, and consequently the laws of God'. Governments must not interfere with 'the great wheel of circulation' even though it dooms 'so many wretches' to 'innumerable servile, degrading, unseemly, unmanly, and often most unwholesome and pestiferous occupations'.

In the 19th century, Bentham relentlessly restated and elaborated Hobbes's atomic individualism, and Benthamism became the dominant ideology. Its doctrine of human nature was summed up in its crudest form in James Mill's article *Government* (1820): 'The desire . . . of that power which is necessary to render the persons and properties of human beings subservient to our pleasures in a grand governing law of human nature.'

So we may say that historically, at least down to 1820 or so, the Palgrave definition is not at all too narrow. But it is too narrow for the latter part of the century, for it leaves out a quite different idea of individualism, one which John Stuart Mill promoted implicitly in his *Principles of Political Economy* (1848) and explicitly in his *On Liberty* (1859), with its opening laudatory quotation from Wilhelm von Humboldt: 'The grand, leading principle, towards which every argument unfolded in these pages directly converges, is the absolute and essential importance of human development in its richest diversity.'

Let us call this *developmental individualism*. It is the antithesis of *possessive individualism*, which assumes that the human being is essentially a striver for, and a receptacle for the acquisition of, material goods. The whole doctrine of *On Liberty* puts Mill squarely in the camp of developmental individualism. And the famous chapter 'Of the Stationary State' (*Principles*, Bk. IV, ch. 6) is eloquent testimony to the depth of his revulsion from the existing acquisitive individualism of the competitive market economy. So, although the developmental ideal of individualism is not found as positively in Mill's *Political Economy* as it is in his *On Liberty*, we may treat the former text also as being on the developmental side. If we do so, however, we must add that Mill was himself so confused a political economist that he did not see that the acquisitive behaviour he denounced was entailed in the capitalist structure he accepted: he did not see that it was that structure which effectively denied a developmental life to the bulk of the wage earners.

A greater political economist than Mill, namely Marx, saw through this confusion and took the logical way out. Marx may be classified as the ultra-collectivist but it is important to see that for him the collective control of the economy was simply a necessary means to an end which was ultra-individualistic, that is, to a flowering of individuality which would be possible when capitalism with its alienation of labour had been surpassed. Marx condemned capitalism morally because it denied any such flowering.

> In bourgeois society . . . the past dominates the present; in Communist society, the present dominates the past. In bourgeois society capital is independent and has individuality, while the living person is dependent and has no individuality. And the abolition of this state of things is called by the bourgeois, abolition of individuality and freedom! And rightly so. The abolition of bourgeois individuality, bourgeois independence, and bourgeois freedom is undoubtedly aimed at (Communist Manifesto, 1848, sect. 2).

And the final outcome of the communist revolution was to be 'an association, in which the free

development of each is the condition for the free development of all' (ibid.).

Similarly:

> In a higher phase of communist society ... after labour has become not only a means of life but life's prime want; after the productive forces have also increased with the all-round development of the individual, and all the springs of co-operative wealth flow more abundantly – only then can the narrow horizon of bourgeois right be crossed in its entirety and society inscribe on its banner: From each according to his ability, to each according to his needs! (Critique of the Gotha Programme, 1875, I, 3).

The *Manifesto's* vision of a fully developed individual as the highest human attainment, echoed in the *Critique of the Gotha Programme*, puts Marx as firmly as Mill in the developmental camp. And just as Mill is there not only by virtue of his *On Liberty* but also by virtue of his *Political Economy*, so Marx is there not only by virtue of the *Manifesto* but also of the *Critique*. And we may add that Marx is there just as firmly in Volume I of *Capital* (1867), where he refers scornfully to the capitalist mode of production as that 'in which the labourer exists to satisfy the needs of self-expansion of existing values instead of, on the contrary, material wealth existing to satisfy the needs of *development on the part of the labourer*' (emphasis added).

There is no warrant in any of this for trying, as some commentators used to do, to drive a wedge between the young 'humanist' Marx and the 'mature' political economist. And, of course, Marx had a strongly developmental vision in his earliest work, the *Economic-Philosophic Manuscripts of 1844*. Thus from his earliest to his latest economic writings there is this development vision. Development individualism is at the very heart of his political economy.

We find, then, that by the time of Mill and Marx developmental individualism is well established: in the liberal tradition it takes place alongside the continuing possessive individualism; in Marx's theory it was inherent from the beginning.

What of the late 20th century? The liberal tradition still contains the two strands of individualism. On the one hand, two of the most esteemed liberal individualists of our time – Isaiah Berlin and John Rawls – are clearly developmental individualists. And on the other hand, the two most noted economic individualists – Friedrich Hayek and Milton Friedman – are equally clearly possessive individualists. Friedman, who would dismantle the welfare state and leave the distribution of economic benefits to an unrestrained competitive market, may be cited as the very model of a possessive individualist. Hayek, whose economic philosophy was set out succinctly in his 1945 lecture *Individualism, True and False*, tries to give market individualism a more agreeable image. He does this by claiming as 'true' individualists the great names in one line of the British tradition, a line from Locke through Mandeville, Hume, Tucker, Ferguson, Smith and Burke, down to Lord Acton, and by categorizing as false individualists the 19th-century Benthamists and Philosophical Radicals, and, on the continent, those infected by Cartesian rationalism, notably the French Encyclopaedists, Rousseau and the Physiocrats. True individualism, he says,

> affirms the value of the family and all the common efforts of the small community and group,... believes in local autonomy and voluntary associations ..., and ... its case rests largely on the contention that much for which the coercive action of the state is usually invoked can be done better by voluntary collaboration.

In sharp contrast, false individualism 'wants to dissolve all these smaller groups into atoms which have no cohesion other than the coercive rules imposed by the state ...'. But Hayek's attempt to humanize market individualism cannot hide the fact that his 'true' individualism, being tied to the free market economy, compels everyone to compete atomistically. Both his kinds of individualism must be graded possessive. Market freedom, the individual freedom to choose between different uses of one's abilities and resources, is, he recognizes, 'incompatible with a full satisfaction of our individual views of distributive justice'. And the individual's freedom is limited by 'the hard discipline of the market'. Hayek's 'true' individualism, for all its smoothness, in the end comes down to

the atomistic 'rugged individualism' of Calvin Coolidge and Herbet Hoover: it is rugged individualism with a smooth false front.

It is clear, then, that the liberal tradition in the late 20th century, including within itself both the developmental individualism of Berlin and Rawls and the possessive individualism of Hayek and Friedman, does contain two antithetical positions and cannot be reduced to either one.

We have said that the old Palgrave definition of individualism was an accurate enough description of the prevailing ideology in the earlier part of the 19th century but was too narrow for the latter part of the century, when the view we have called developmental individualism emerged alongside of the earlier purely possessive individualism. We may go on to ask, what brought about this change? What brought developmental individualism into the picture?

Clues are to be found in John Stuart Mill's own writings. In the first place is his perception that the unrestrained market economy had produced a kind of society which would no longer be tolerated by the working class it had produced. In his 1845 article 'The Claims of Labour' he took the rise of the Chartist movement, with its threat of physical force, to be evidence that the British working classes would no longer put up with things as they were, and he believed that 'the more fortunate classes' must see the writing on the wall: 'While some, by the physical and moral circumstances which they saw around them, were made to feel that the condition of the labouring classes *ought* to be attended to, others were made to see that it *would* be attended to, whether they wished to be blind to it or not.'

In the second place, perhaps partly because of this apprehension of class violence, Mill became a more sensitive and humane liberal than his father or Bentham, denouncing as utterly unjust the existing relation of effort and reward, by which the produce of labour was apportioned 'almost in an inverse ratio to the labour' (*Principles of Political Economy*, Bk. II, ch. 1, sect. 3), and deploring the fiercely competitive character of the market-dominated society of his day, 'the trampling, crushing, elbowing, and treading on each other's heels, which form the existing type of social life' (*Principles*, Bk. IV, ch. 6, sect. 2). In reacting as early as 1848 against this kind of society, Mill was a harbinger of the more humane social conscience which became noticeable in early 20th-century liberal thinking and which in mid-20th century brought the welfare state.

We conclude that the old Palgrave definition of individualism, already too narrow when it was promulgated, became increasingly inadequate in the subsequent decades. It was made inadequate by the rise and growth of developmental individualism, which in turn was the result of two distinct but related phenomena – the apprehension by middle-class thinkers of a danger of working-class violence, and the somewhat delayed reaction of those same minds to the shocking brutality of the industrial *laissez-faire* society. The two factors together ensured that developmental individualism would coexist with possessive individualism in the heyday of free capitalist enterprise.

How much longer they will coexist is not readily predictable. The danger of class violence now within advanced capitalist welfare states is less than Mill thought it to be in the society of his time, but what may well be called class violence as between undeveloped (or misdeveloped) and developed states is not far to seek in our time. And the working and living conditions of wageearners in developed countries are less savage now than they were in Mill's day, but the increasing speed and tension of much of the work presses heavily on them. All we can say is that the probability of our advanced societies continuing to afford any substantial measure of developmental individualism varies inversely with the degree of industrial speed-up and the amount of class violence, national and international.

See Also

- ▶ Altruism
- ▶ Economic Man
- ▶ Self-interest

Bibliography

Burke, E. 1795. *Thoughts and details on scarcity, originally presented to the Right Hon. William Pitt, in the month of November, 1795*. London, 1800.
Hayek, F.A. 1946. *Individualism: True and false. The twelfth Finlay Lecture. . . . 1945*. Dublin: Hodges, Figgis & Co.; Oxford: B.H. Blackwell.
Lukes, S. 1973. *Individualism*. Oxford: Blackwell.
Macpherson, C.B. 1962. *The political theory of possessive individualism*. Oxford: Oxford University Press.
Marx, K. 1867. *Capital*. London: Lawrence & Wishart, 1970.
Marx, K. 1891. *Critique of the Gotha Programme*. London: Lawrence & Wishart, 1938.
Marx, K. 1959. *Economic and philosophic manuscripts of 1844*. Moscow: Foreign Languages Publishing House.
Marx, K., and Engels, F. 1848. *The communist manifesto*. London.
Mill, J. 1820. *Government*. (Originally written for the supplement to the fifth edn of the *Encyclopaedia Britannica* which was completed in 1824.)
Mill, J.S. 1845. The claims of labour. *Edinburgh Review*. In *Collected works of John Stuart Mill*, Vol. IV, ed. J.-M. Robson, 363–389. Toronto: University of Toronto Press, 1967.
Mill, J.S. 1848. *Principles of political economy*. 2 vols, London: J.W. Parker.
Mill, J.S. 1851. Newman's political economy. *Westminster Review*. In *Collected works of John Stuart Mill*, Vol. V, ed. J.M. Robson, 439–457. Toronto: University of Toronto Press, 1967.
Mill, J.S. 1859. *On Liberty*. London: J.W. Parker.
Mill, J.S. 1879. Chapters on socialism. *Fortnightly Review*. In *Collected works of John Stuart Mill*, Vol. V, ed. J.-M. Robson, 703–753. Toronto: University of Toronto Press, 1967.
Palgrave, R.H.I. (ed.) 1894–1899. *Dictionary of political economy*. London: Macmillan & Co.

Individualism Versus Holism

Harold Kincaid

Abstract

Issues about individualism and holism in economics surface because economics is committed to understanding both institutions and large-scale economic processes, in terms of constrained maximizing of individuals. Three key questions are at issue. Can a theory of individual economic behaviour capture everything we want to explain about the economy in principle? To what extent do our accounts of individual economic behaviour trump or constrain other economic explanations that are not directly about individuals? Are non-individual economic entities real, and what is their relation to individual behaviour? These questions are answered in light of developments in economics and in philosophy of science.

Keywords

Causation; Evolutionary game theory; Explanation; Game theory; Holism; Hoover, K.; Methodological individualism; New institutional economics; Rational expectations; Reductionism; Representative agents; Schumpeter, J. A.; Smith, A.

JEL Classifications

B4

The idea that economic outcomes result from and thus are to be explained by the maximizing choices of individual human beings has been essential to economics at least since Adam Smith. Yet this maxim of methodological individualism has consistently existed alongside and in tension with the important role that institutions play in economic outcomes and the desire of economists to explain large-scale phenomena such as the rate of inflation and unemployment. Debates over individualism and holism have generally been vaguely formulated and argued at an abstract level with a questionable relationship to the actual practice of economics. The purpose of this article is to clarify the theses and arguments at work and replace rhetoric with identifiable empirical issues with real ties to economic practice. Some recent developments in economics – for example, the new economics of information asymmetry and institutions, and the obstacles to the refinement programme in classical rational choice game theory and to rational expectations programmes in macroeconomics – argue against

the most extreme forms of individualism while leaving a plausible place for various more modest individualist constraints.

There are at least three questions at the centre of economic debates over holism and individualism:

1. Can a theory of individual economic behaviour capture everything we want to explain about the economy in principle?
2. To what extent do our accounts of individual economic behaviour trump or constrain other economic explanations that are not directly about individuals?
3. Are non-individual economic entities real, and what is their relation to individual behaviour?

The first question can be thought of as a question about theory reduction. Can a well-formulated theory of individual behaviour replace all economic explanations that are not directly about individuals, at least those we think are relatively well confirmed? The second question is usually put as a thesis about mechanism: every economic explanation has to be given individualist mechanisms. The final question is about ontology: what entities populate the economic realm and how are they related?

Individualists tend to answer the first two questions affirmatively and either deny that social entities exist at all or assert that individuals are in some sense prior to them. Extreme holists take the opposite stance, answering negatively to the first two questions and arguing that social entities are real and in some sense prior to individuals.

These three questions are no doubt related, and it is often asserted that an answer to one of these questions tells us the answer to the others. Yet, for the most part, discussions in the literature do not clearly identify which theses are at issue nor exactly what their relationship is to each other.

To what extent is economics committed to a version of individualism? Schumpeter (1954), in his classic history of economic analysis, apparently coined the term 'methodological individualism' and argued that it, along with a fundamental focus on prices and general equilibrium analysis, was the common core of economics since Smith. There is little evidence that Schumpeter was right about the classical economists. They were interested in the distribution of the total economic product to social classes and the factors influencing its growth. Their accounts frequently involved taking institutional structure as given rather than explained; Smith explicitly acknowledged that invisible hand processes work against a background of social institutions, customs, and the like (Gordon 1991).

However, the neoclassical revolution certainly ushered in an explicit commitment to individualism. Many past and recent elements of modern economics are directly motivated by the individualist theses mentioned above, among them:

(a) the general equilibrium programme of explaining all economic phenomena on the basis of individual preferences and initial endowments;
(b) the rational choice game-theory programme of explaining norms, institutions, the behaviour of the firm, and so on, completely in terms of the behaviour of maximizing individuals; and
(c) the rational-expectations programme which seeks to model macroeconomic phenomena in terms of the expectations states of individual maximizing agents with given preferences, technology, and so on.

These doctrines make the three kinds of claims listed above: (*a*) the ontological claim that all economic phenomena consist of the actions of individuals, (*b*) the reductive claim that a theory based on such primitives can explain and thus reduce all economic phenomena to individual behaviour, and (*c*) the claim that individual-based mechanisms are a requirement of good explanation. Thus the individualism–holism controversy in economics crucially involves, at a minimum, evaluating the extent to which these programmes succeed in realizing their individualism, putting aside other worries such as the assumption of equilibrium outcomes and so on.

To start with the ontological issues first, there is no good evidence that I know of that any major economist from the classicals on affirmed the

extreme holist ontological thesis that society acts or exists entirely independently of the behaviour of individuals. Moreover, the question whether aggregate economic entities are real seems to me to be the least interesting and controversial issue in the individualism–holism debate in economics. No one denies that firms, for example, are collections of individuals. If each of the individuals in the firm exists, then the sum of them exists as well. The real issue I would argue is how far we can go in *explaining* the aggregate in terms of the individuals composing it – that is a question of reduction.

Hoover (2001) has argued for the reality of macroeconomic aggregates on the grounds that something is real if it stands in causal relations and that macroeconomic aggregates do stand in such relations. However, the advocate of rational expectations is not denying that the GDP or the rate of inflation exists but instead is asserting that their causal efficacy can be explained in terms of the actions of individuals and that, given rational expectations, we cannot expect there to be stable causal relations among aggregates invariant to policy changes. Again the issues seem to be more about explanation, not whether aggregates are real.

The second ontological claim – that the characteristics of economic aggregates are dependent upon the facts about individuals – has more content. Borrowing from philosophical discussions of physicalism (Hellman and Thompson 1975), reductionism in general and in the philosophy of mind in particular (Fodor 1974), we can take this claim to be asserting that all non-individual facts supervene on or are determined by the facts about individuals. The basic idea is that one set of facts A supervenes on or is determined by another set B just in case once the B facts are set, so are the A facts. In other words, there is no difference in the A facts without a corresponding difference in the B facts. As we will see below, this asserts only a one-way conditional from the Bs to the As and not the stronger biconditional of A if and only if B that is typical of reduction. (So the individualist thesis would be that the economic facts about individuals fix the facts about other aggregate or collective economic entities.)

Talk of 'facts' is vague. We can be more precise by asking if the truths or assertions of a particular theory fix those of another – in this case, whether a particular economic theory referring only to individuals determines or fixes the truths of an economic theory that includes terms referring to collective economic phenomena. Put this way, the debate over this individualist thesis is really many different debates, depending on what particular models are at issue, and individualism might be plausible in some cases and not others. So, for example, we can ask whether downward-sloping demand curves of consumer choice theory ensure that aggregate market demand curves are likewise downward-sloping. The evidence seems to be that they are only under very restrictive conditions that are unlikely to hold (Deaton and Muellbauer 1980). More generally, the Sonnenschein–Mantel–Debreu theorem shows that individual excess demand functions do not ensure a unique equilibrium. Similar difficulties face some attempts to derive macroeconomic implications from choice theoretic models of individual behaviour (Martel 1996). Yet it seems clear that there must be *some* model of individual behaviour that determines such aggregate relations.

Individualism acquires it clearest statement as a claim about theory reduction. There is an extensive literature on the requirements for theory reduction in general as exemplified by the reduction of the gas laws to statistical mechanics. The gas laws refer essentially to temperature while no such notion is a fundamental category in statistical mechanics. However, temperature has an analogue in the mean kinetic energy of the molecules in a gas – we can define the former in terms of the latter. A definition requires at a minimum some kind of biconditional relationship between the terms involved. When we can produce such a definition, then to reduce we should be able to reproduce the explanation given by one theory in the vocabulary of another, for example, by showing that the gas laws follow from laws of statistical mechanics once temperature is equated with mean kinetic energy.

There are at least three possible ways that one theory might turn out to be irreducible (Kincaid 1996, 1997).

I

Multiple realizations: if we wanted to reduce ordinary claims about chairs, for example, to particle physics, then we would need to find a one-to-one correspondence between chair categories and quantum mechanical descriptions, because there are indefinitely many ways to bring chairs about in physical terms and no natural way to capture them in terms of physics. Chairs are in that sense multiply realized and thus there is no link that allows physical explanations to replace common sense ones. The root idea here is thus that categories at one level of description may pick out kinds that look disparate in another vocabulary. It is important to note that the multiple realizations problem undermines the common conclusion that aggregate economic phenomena *must* be reducible because they are made from individual behaviour. Reduction is a claim about what specific theories can in principle explain. From the fact that As are composed of Bs it does not follow that a specific theory applying to the Bs has the explanatory resources to eliminate explanations in the categories that describe the As – we make various claims about chairs that cannot be cashed out in quantum mechanics even though chairs are made entirely of atoms.

One–many relations: reductive definitions can fail in the other direction in that in the reducing theory the descriptions used are not sufficient to fix the descriptions in the theory to be reduced. This is failure of the one-to-one mapping as well, but in the other direction.

Presuppositions: the reducing theory may find itself implicitly in need of categories from the theory to be reduced in its own accounts. To take an example from reductionism debates outside economics, attempts to explain antibodies in purely biochemical structural terms arguably fail, because in the end none of the physical descriptions suffices unless an immune response also occurs (Kincaid 1997). But appealing to immune response seems not be giving a physical explanation but a biological one.

It is an empirical issue whether in fact these sorts of obstacles are real for reduction in economics. When and where they are real need not be uniform across every economic sub-domain and economic model – reduction might be feasible for some and not for others. However, there is some good evidence to think that reduction is often unfeasible for the following kinds of reasons:

1. Much explanation in economics that might seem individualist in spirit is really nothing of the sort. One case in point is the widespread use of representative agents who are not flesh and blood individuals and who cannot be legitimized as reasonable aggregations of individual behaviour. Another is the widespread practice of taking household and firms as basic entities. These are social, aggregative entities that, when treated as black boxes, belie a commitment to individualism.
2. There are various reasons to think that multiple realizations of economic categories in terms of individual behaviour are likely. (a) Many claims in economics are at least implicitly motivated by selectionist arguments, for example, that firms must be profit maximizers if they are not to be weeded out by competitive selection. However, such selective mechanisms do not 'care' about how profitability comes about, only that it does. This means that there may be multiple ways of organizing individual behaviour that meet the criterion. This possibility is reinforced by results in the theory of the firm, where there are many plausible models of how profit- maximizing behaviour might be brought about by the organization of individual incentives. (b) Much applied economics consists of estimating aggregate supply and demand curves in specific markets. This work proceeds quite well without estimating individual utility functions. This suggests that these aggregate phenomena are indifferent to the exact details of individual behaviour. Becker's (1976) argument that downward-sloping demand curves would be expected from random choices given budget constraints provides a theoretical account of why this is likely to be the case. (c) Solid results from physics and elsewhere suggest that complex causality in aggregate phenomena often show structural relations or 'universalities' (Batterman 2001) that are indifferent to a wide range of underlying detail and that in fact are described by categories that are

'scale relative' in that they have no counterpart in smaller scales of resolution (Ladyman and Ross 2007). Hoover (2001) makes an argument like this about macroeconomic variables: the rate of inflation or the GDP has no obvious meaning at very fine scales of measurement. Similarly, equilibrium analysis in terms of strong attractors and the like in evolutionary game theory also presents a parallel situation. The properties of an equilibrium can be understood while there is a wide range of actual dynamic paths to that equilibrium, the details of which are inessential to the equilibrium explanation.

3. Economic explanations involving individuals often rest on – they take as given and unexplained – information about institutions, structures, norms, and so on that are not cashed out in individualist terms. In short, they presuppose rather than eliminate social processes. As we noted above, Schumpeter's claim that economics since its inception has been methodological individualist in orientation is implausible in any strong form, because Smith, for example, is quite clear that institutions and customs matter in fundamental ways. Not surprisingly much work in economics after the neoclassical revolution has carried the banner of methodological individualism in its rhetoric but its practice is much closer to Smith.

One clear illustration of this comes from recent developments in rational choice game theory explanations in economics. The failure of the refinement programme to plausibly eliminate all multiple equilibria means that the focal points that are often used to explain which equilibrium is selected will bring in unexplained norms. Bayesian agents in games reach equilibrium when they have sufficiently similar priors, assuming rather than explaining the social processes that produce consensus (Janssen 1993). Most fundamentally, game theory explanations have to take as given the possible payoffs, the utility functions of individuals, the information available to them, and the initial distribution of resources. This assumes rather than explains much institutional structure. Property rights have to be defined and so on. Much of the new institutional economics is about how these institutional differences can have strong influences on outcomes. The conclusion to draw is that explanations in terms of individuals have to be supplemented with accounts of collective social and economic phenomena, making reduction – full explanation in individual terms – unlikely.

Another example where the individualist rhetoric can outrun the actual practice comes from explanations of the distribution of income in standard neoclassical models. The goal is to explain what individuals get in terms of the traits of individuals, for example, investment in human capital and so on. Institutional structure is in the background here as well. The preferences of workers and initial distributions of wealth are taken as given. It is generally assumed that there is a direct link between productivity and earnings, which the considerable work on the theory of the firm shows holds only under specific institutional contexts that are often not satisfied. Most fundamental, those models generally take the distribution of jobs or positions as given. In effect, what is being explained concerns what determines on which rung on the ladder individuals stand. Left unexplained is the number of rungs and the distances between them (see Sattinger 1993).

Of course, nothing precludes the individualist from seeking further explanations of all these unexplained collective phenomena in purely individual terms. But the breath of these problems and the lack of individualist explanations at this point suggest that the current evidence for the reductionist version of individualism in economics is slim.

I turn finally to the version of individualism claiming that individualist *mechanisms* are necessary. Here I think is the most important individualist insight, though it is important to distinguish various versions of this claim, for some are considerably more plausible than others. The notion of a mechanism is nebulous. The root idea, going back to Maxwell and before in physics, seems to be that of a continuous causal process – one that is not gappy as it were. Taken that way, a mechanism might be either horizontal or vertical. To find the causes between A and B is to find a horizontal

mechanism and explaining how the parts of A contribute to its causal influence on B is identifying the vertical sense of mechanism. A related important distinction concerns how a mechanism is described and in what detail. An 'antibody' and a 'compound of such and such a structure' may commit us to different things.

With these distinctions in hand, here are some general things that can be said about mechanisms in science in general. We can sometimes know that A causes B without knowing either the horizontal or the vertical mechanism. To cite a common sense example, I can know that the flying baseball caused the broken window without knowing the quantum descriptions of the baseball's constitution or the exact details of how the ball surface interacted with the glass. Furthermore, the notion of knowing the 'full' mechanism is not well defined, since we can generally give more fine-grained descriptions of the constituting parts of or of the time periods between causes; no account of a causal process is a complete explanation in the sense that there are no unanswered questioned that might be answered. Finally, the place of specific mechanisms in our accounts of the world seems to depend on three things: how solid our knowledge is at the level of description we are using to pick out the mechanism, how solid our knowledge is about that process for which we are seeking mechanisms, and to what extent the two make presuppositions about the other. An account of large-scale brain structures, for example, that required neuronal processes at speeds beyond the known synaptic firing times would be suspect. An explanation that was well confirmed at the scale of brain structures through experiment and physical tracing and that relied on no very specific view of neuronal details should not be strongly constrained by molecular mechanism, particularly if our understanding of the molecular details was much less solid than our understanding at the level of brain structure.

From this general perspective, some claims that individualist mechanisms are essential in economics are plausible and some are not. Among the implausible is that no economic explanation ever succeeds until there is an account in terms of individual maximizing behaviour and general equilibrium. There is just too much good work in economics that provides apparently well-confirmed explanations without meeting this requirement. As noted above, much applied economics is about aggregate supply and demand that has no general equilibrium foundations. Other compelling explanations in industrial organization describe firm behaviour in competitive environments of various kinds with no pretence of providing a foundation in individual (as opposed to firm) maximizing behaviour. Good econometric work in macroeconomics can use structure breaks between macroeconomic variables to show causation without any account of underlying individual behaviour (see Hoover 2001).

Of course, these accounts could certainly be made stronger by providing some account of how they relate to individual behaviour. Given everything we know from experimental and behavioural economics, however, the theory would not be a simple picture of individuals maximizing utility functions. In any case, the fact that nonindividualist explanations can be made stronger does not thereby mean they are bad explanations – if does not follow from the fact that I cannot answer all questions about a domain that I can answer none.

Alternatively, mechanisms in terms of individual behaviour can be plausible requirements indeed in the right circumstances. Critics of Keynesian orthodoxy had reason to be critical in that Keynesian models required individuals to be systematically fooled. Critics of rational expectations models, however, could with equal justification turn the tables and reject those models because of their lack of individualist mechanisms in that they require individuals to make the best econometric forecast given the available data. This thus illustrates the theme of this article, which is that, once we move beyond the individualist rhetoric typical of the economics profession, individualism and holism have many different claims that vary enormously in plausibility – the devil is in the details.

See Also

- ▶ Aggregation (Theory)
- ▶ Explanation
- ▶ Methodological Individualism

Bibliography

Batterman, R. 2001. *Devil in the details*. Oxford: Oxford University Press.
Becker, G. 1976. *The economic approach to human behavior*. Chicago: University of Chicago Press.
Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behavior*. Cambridge: Cambridge University Press.
Fodor, J. 1974. Special sciences (or the disunity of science as a working hypothesis). *Synthèse* 28: 97–115.
Gordon, S. *1991. The history and philosophy of social science. London*: *Routledge*.
Hellman, G., and F. Thompson. 1975. Physicalism: Ontology, determination, and reduction. *Journal of Philosophy* 72: 551–564.
Hoover, K. 2001. *Causality in macroeconomics*. Cambridge: Cambridge University Press.
Janssen, M. 1993. *Microfoundations: A critical inquiry*. London: Routledge.
Kincaid, H. 1996. *Philosophical foundations of the social sciences*. Cambridge: Cambridge University Press.
———. 1997. *Individualism and the unity of science*. Lanham: Rowman and Littlefield.
Ladyman, J., and D. Ross. 2007. *Every thing must go*. Oxford: Oxford University Press.
Martel, R. 1996. Heterogeneity, aggregation, and a meaningful macroeconomics. In *Beyond microfoundations*, ed. D. Colander. Cambridge: Cambridge University Press.
Sattinger, M. 1993. Assignment models of the distribution of income. *Journal of Economic Literature* 31: 831–880.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Indivisibilities

William J. Baumol

Abstract

A commodity is indivisible if it has a minimum size below which it is unavailable, at least without significant qualitative change. Indivisible inputs yield economies of scale and scope. But even where indivisibilities impose large fixed costs, if they are not sunk, potential competition can impose behaviour upon incumbents that is consistent with economic efficiency. Perhaps the most significant way in which indivisibilities can impede efficiency in pricing is the existence of indivisible input–output vectors that are efficient but which are not profit maximizing at any positive scalar prices. Integer programming is naturally suited to optimality analysis involving indivisibilities.

Keywords

Barriers to entry; Circular flow; Economies of scale; Economies of scope; Euler's theorem; Fixed costs; Indivisibilities; Integer programming; Marginal cost pricing; Natural monopoly; Non-convexity; Ramsey pricing; Sunk costs

JEL Classifications

D2

A commodity is indivisible if it has a minimum size below which it is unavailable, at least without significant qualitative change. Most commodities are indivisible but this is often unimportant. Half a chair has little use, but this makes little difference for analysis of market demand because so many are sold that there is little inaccuracy in treating an increase in sales of chairs from 10 million to 10,000,001 as a change in a continuous variable. In other cases, minimum size is so large relative to usage that it requires special analytic approaches and has substantial behavioural consequences: a Boeing 747 passenger aircraft is a large outlay for any airline; to carry *any* freight from New York to Chicago a railroad must lay at least two rails, each about 1,000 miles long.

Fixed Cost and Sunk Cost

The *fixed cost* of a firm is defined as the minimum outlay it must incur to carry out any activity. If we write (assuming input prices fixed) the long-run

cost function as $C(y) = k + f(y)$, where $k =$ constant, $f(0) = 0$, and $y =$ the vector of output quantities, then k is the fixed cost. As in the railroad example, the need for indivisible equipment is the normal source of fixed costs.

Fixed costs are important in economics as a source of economies of scale, of impediments to the workings of the price mechanism, of breakdown in the convexity conditions usually relied upon in optimization calculations and in the uniqueness of solutions.

Fixed costs are often confused with *sunk costs*, which are also related to indivisibilities. A sunk cost may or may not be larger than the minimum outlay a firm needs to operate but, once incurred, it cannot be withdrawn for some substantial period without significant loss. An automobile producer may build a plant much larger than the minimum needed to turn out one car, and once the capital is sunk it may only be possible to retrieve it gradually as vehicles are sold. Thus, sunk costs (like the car plant) need not be fixed and fixed costs (like an aircraft) need not be sunk.

Economies of Scale and Scope

Indivisible inputs by their nature yield economies of scale and scope. An indivisibility requires a producer of even a small output volume to acquire relatively large capacity, part of which must be unused. The firm can then increase its outputs without increasing costs proportionately (economies of scale). Formally, strict economies of scale are defined to be present at output vector y if $C(ay)/a > C(y)$, where $0 < a < 1$, that is, if average cost is declining along the ray ay. With fixed costs this becomes $[k + f(ay)]/a > k + f(y)$, $k > 0$. Then, assuming that $f(y)$ is bounded from both above and below, say, $0 \leq f(y) \leq M < \infty$, the scale economies criterion must clearly be satisfied as a approaches zero. Thus, the presence of fixed costs always introduces scale economies (so defined), at least in any neighbourhood of the origin.

If the indivisible item is not too specialized, the firm can add commodities to its product line without the combined costs equalling the sum of those of several more specialized enterprises which together produce the same output vector as our firm. The latter attribute is referred to as *economies of scope*. Formally, using the three product case $y = (y_1, y_2, y_3)$ for simplicity, strict economies of scope are defined by $C(y) < C(y_1, 0, 0) + C(0, y_2, 0) + C(0, 0, y_3)$.

Together, economies of scale and scope are what underlie the phenomenon of natural monopoly. An industry is said to be a *natural monopoly* at output y if one single firm can produce y more cheaply than can be done by any combination of two or more firms. Formally, if y^i is the output vector of firm i, then the industry is a natural monopoly at y if $C(y) < \Sigma\ C(y^i)$ for each and every set of y^i such that $\Sigma\ y^i = y$.

Scale economies lead to natural monopoly because in their absence it may be possible to save resources by dividing the industry's output among several firms, each providing similar proportions of the industry's output vector. Specifically, the absence of (weak) scale economies at y means that, for some values of a, $C(ay)/a < C(y)$, $0 < a < 1$. Suppose there exists such a value of a at which $b = 1/a$ is an integer. Then the industry can reduce cost by dividing output among b firms each producing $y^i = ay$, at total cost

$$\sum C(y^i) = bC(ay) = C(ay)/a < C(y),$$

thus violating the criterion of natural monopoly. Economies of scope are relevant because in their absence it may be possible to save resources by dividing up theindustry's products among specialized enterprises. Specifically, for example in the two-product case, absence of weak economies of scope means $C(y_1, 0) + C(0, y_2) < C(y) = C(y_1, y_2)$, also violating the natural monopoly requirement.

It can also be shown that scale economies together with an attribute closely related to economies of scope are sufficient (but not necessary)

for an industry to be a natural monopoly (see Baumol et al. 1982, pp. 178, 187–8).

Indivisibilities, Sunk Costs and Barriers to Entry

The literature offers various definitions of 'barriers to entry', some mutually inconsistent. If one defines them as impediments to the invisible hand mechanism, then sunk costs are entry barriers while fixed costs are not.

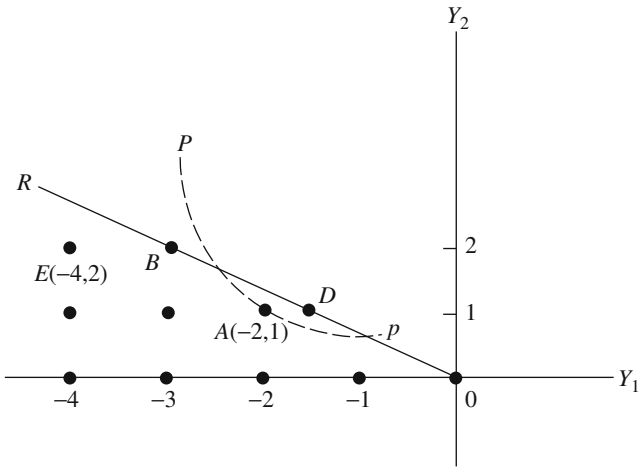
The need to sink capital into an enterprise constitutes a risk which obviously can deter a potential entrant and thus can protect incumbents from potential competition. So, in an industry with relatively large sunk costs, monopoly profits and inefficiencies become possible.

On the other hand, even where indivisibilities impose large fixed costs, if they are not sunk, potential competition can impose behaviour upon incumbents that is consistent with economic efficiency. Where the fixed capital is highly mobile and there is an active market on which it can readily be sold (as with, for example, ocean cargo vessels) then the fixed capital constitutes no special risk and is no impediment to entry. Even if the indivisibilities make the industry a natural monopoly it will be unable to earn excess profits, operate inefficiently or behave like a protected monopolist in other ways, because this will attract entry that – with no sunk costs – incurs little risk and punishes the misbehaving monopolist.

Indivisibilities as Impediment to Efficient Pricing

Perhaps the most significant of the ways in which indivisibilities can impede efficiency in pricing is the existence of indivisible input–output vectors that are efficient but which are not profit maximizing at any positive scalar prices. This is best shown diagrammatically. In Fig. 1 (Frank 1969, pp. 5, 42–3), $y_1 \leq 0$ and $y_2 \geq 0$ are the input and ouput quantities respectively. With both of them indivisible, the dots, or *lattice points*, represent the only feasible input–output combinations. Point $A = (-2, 1)$ is efficient since no feasible lattice point lies to its northeast. However, A lies inside the convex hull of the (non-convex) feasible region whose northeast boundary is ray OR. Hence, any line given by $p_1y_1 + p_2y_2 =$ profit, through point A must lie below at least one lattice point on OR (here, either B or 0). Thus, at any non-negative prices efficient point A must be less profitable than 0 or B – no simple prices can lead profit maximizing firms to produce A. Only a set of

Indivisibilities, Fig. 1



'nonlinear prices' (for example, two-part tariffs), which lead to a curved isoprofit locus such as *PP*, can induce production of *A*.

The diagram demonstrates how indivisibilities lead to non-convexity. For example, a line segment connecting points *A* and *B* in Fig. 1 clearly is not composed entirely of lattice points, that is, it is not entirely contained in the feasible set of lattice points, and so that set is not convex.

The graph also shows in another way how indivisibilities introduce scale economies. *C*onsider *D*, a non-lattice point on *OR* to the right of *A*. Let c be the smallest integer for which c(distance AD) ≥ 1 (in the graph $c = 2$). Then cA will be a feasible lattice point (point *E*), but there will be a point (*B*) between *E* and $E - c(AD)$ which is a feasible lattice point, with the same output and a smaller input quantity than those at *E*. Since *E* is an integer multiple of efficient point *A*, one can multiply output by $c > 1$ while multiplying input by a smaller amount, that is, there must be scale economies.

Indivisibilities impede the price system in yet another way. By creating scale economies they make marginal cost pricing unprofitable. Specifically, let y be an output vector at which there are scale economies so that $C(ay) = a^b C(y)$ in the neighbourhood of y, with $b < 1$. Then, the function is locally (approximately) homogeneous of degree b and by Euler's theorem $\Sigma\, y_i \partial\, C/\partial y_i = bC < C$. Hence, if prices are set equal to marginal costs the supplier must lose money. In that case, financial feasibility requires the substitution of Ramsey prices (see Ramsey pricing) for marginal cost prices to achieve a second-best optimal resource allocation. This is true not only for the individual firm – the entire economy may have no parametric price option that is superior to Ramsey prices. For all outputs must be sold to suppliers of inputs and the receipts from output sales are paid out as wages, profits, and so on, to the input suppliers. This imposes (in the absence of lump sum payments with parametric prices of inputs and outputs) the economy's circular flow requirement $\Sigma\, p_i y_i = 0$, again taking input quantities to be negative. Now, a set of Pareto will, in optimal prices p_i^* general, not satisfy this constraint. The second-best prices, p_i, which are constrained to satisfy this requirement, are by definition the Ramsey prices and the differences $t = p_i - p_i^*$ between Ramsey prices and optimal prices may be interpreted as the optimal vector of taxes needed for compliance with the economy's circular flow constraint.

That is the form in which Frank Ramsey's original treatment is expressed. As we have just seen from the Euler's theorem argument, where costs are differentiable with respect to outputs, the first-best prices of the outputs, which are their marginalcosts, will not satisfy the circular flow constraint when there are scale economies. This shows that in general, where indivisibilities create scale economies, optimality in pricing cannot avoid the complications of Ramsey theory.

There is a third way in which indivisibilities complicate the optimization process. As is well known, where the feasible set is not convex, as must be true when there are indivisibilities, a multiplicity of local maxima is likely to be present and an iterative solution process that always follows a direction in which profit (or the value of the social objective function) is increasing may well lead towards a local optimum rather than one which is global.

Integer Programming and the Analysis of Indivisibilities

Integer programming is the mathematical technique that is naturally suited to optimality analysis involving indivisibilities. An integer programme is a mathematical programme in which only integer values are admissible for some or all of the variables. The constraint requiring $x =$ number of locomotives to be an integer is what keeps the solution from including the absurd recommendation that 1.783 locomotives be produced.

Integer programming also permits the solution of more subtle indivisibility problems, such as those involving scale economies or either/or choices, which have resisted other analytical techniques. As an example, consider a firm required to produce y units of output using either a machine of type 1 or a machine of type 2, where x is the

vector of other inputs, and x_1 and x_2 are the respective numbers of the two types of machines purchased, $\Pi(y, x, x_1, x_2)$ is the profit function and $y \leq f(x, x_1, x_2)$ is the production constraint. Then the firm must

$$\text{maximize } \Pi\,(y, x, x_1, x_2)$$

subject to the constraints

$$\begin{aligned} y &\leq f(x,\ x_1,\ x_2) \\ y, x, x_1, x_2 &\geq 0 \\ x_1 + x_2 &\leq 1 \\ x_1, x_2 &\text{ integer.} \end{aligned}$$

The last two constraints guarantee that x_1 will take either the value zero or unity and that (at least) one of them will be zero, as an either/or decision requires.

Economies of scale and scope raise related issues. Such cases tend to yield corner rather than interior solutions. If there are n firms, each with different attributes, which are candidate producers of industry output vector y, it is likely to be most economical for just one of them to produce all of y. But which one of the n firms should do the job? That is obviously an extended either/or issue whose formal statement is perfectly analogous to that just described.

Indivisibilities give rise to other complex combinatorial problems. The choice among m machines may, for example be constrained by the fact that a machine of type A will work only if a machine of type B is also purchased. This is dealt with via the constraints $x_a \leq x_b$, x_a, x_b integer. In such problems the indivisibility feature is fundamental and cannot be avoided by non-integer approximation. In sum, indivisibilities raise basic issues for theory and for methods of analysis which bear little resemblance to those pertinent to cases of divisibility.

See Also

- ► Contestable Markets
- ► Ramsey Pricing

Bibliography

Baumol, W.J., J.C. Panzar, and R.D. Willig. 1982. *Contestable markets and the theory of industry structure*. San Diego: Harcourt Brace, Jovanovich.

Dupuit, J. 1844. De la mésure de l'utilité des travaux publiques. *Annales des Ponts et Chausées*, 2nd Series, vol. 8. Reprinted in *International economic papers* no. 2 (1952). London: Macmillan.

Frank, C.R. Jr. 1969. *Production theory and indivisible commodities*. Princeton: Princeton University Press.

Gomory, R.E. 1965. On the relation between integer and non-integer solutions to linear programs. *Proceedings of the National Academy of Sciences* 53: 260–265.

Gomory, R.E., and W.J. Baumol. 1960. Integer programming and pricing. *Econometrica* 28: 521–550.

Koopmans, T.C. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.

Lewis, W. 1949. *Overhead cost*. London: George Allen & Unwin.

I

Induction

Paul W. Humphreys

Induction, in its most general form, is the making of inferences from the observed to the unobserved. Thus, inferences from the past to the future, from a sample to the population, from data to an hypothesis, and from observed effects to unobserved causes are all aspects of induction, as are arguments from analogy. A successful account of induction is required for a satisfactory theory of causality, scientific laws, and predictive applications of economic theory. But induction is a dangerous thing, and especially so for those who lean towards empiricism, the view that only experience can serve as the grounds for genuine knowledge. Because induction, by its very nature, goes beyond the observed, its use is inevitably difficult to justify for the empiricist. In addition, inductive inferences differ from deductive inferences in three crucial respects. First, the conclusion of an inductive inference does not follow with certainty from the premises, but only with some degree of probability. Second, whereas valid deductive inferences retain their validity when

extra information is added to the premises, inductive inferences may be seriously weakened. Third, whereas there is widespread agreement upon the correct characterization of deductive validity, there is widespread disagreement about what constitutes a correct inductive argument, and indeed whether induction is a legitimate part of science at all.

Approaches to these issues generally fall into two categories. The older, more philosophical approaches attempt to provide an extremely general justification for the use of inductive methods and to isolate the universal characteristics which make for a correct inductive inference. The second kind of approach focuses on what are called 'local inductions' – analyses of very specific kinds of inferences, applicable in precisely detailed circumstances. With the enormously increased power of statistical methods which is characteristic of this century, the second approach has become more and more the province of theoretical and applied statistics. It would be inappropriate to discuss specifically statistical issues here and the reader is referred to the excellent Barnett (1982). However, it should be recognized that although these detailed mathematical techniques have increased our understanding of induction immensely, they do not by themselves answer all questions about the soundness of inductive procedures. In collecting data, for example, judgements must be made about which situations are similar to one another, and hence a combination of analogical principles and judgements of causal relevance will be needed. The principles of experimental design, not only for field data but also in the growing subject of experimental economics, generally require such judgements. Bayesian statistical methods need principles on which to attribute prior probabilities, and there is an extensive philosophical literature on the acceptability of such principles. Finally, it should be emphasized that most statistical techniques were developed within a climate of extreme empiricism or positivism, and that the application and integration of statistical models to economic systems requires a delicate inductive sensibility that cannot be reduced to algorithmic procedures.

Philosophical Approaches

The use of induction as the basis of a general scientific method was first systematically advocated by Francis Bacon. His suggested methods will seem queer to the modern reader, but the importance of his break with the deductive traditions of Greek and medieval thought should not be undervalued. He himself realized this in entitling his principal work *Novum Organum* to mirror Aristotle's *Organum* of logic, and his methods partially anticipate the eliminative methods later championed by J.S. Mill and Popper. Important as Bacon's work was, all modern work on induction lives under the shadow of the later 'problem of induction'.

The problem of induction' is to state conditions under which an inductive inference can be rationally justified. Ever since its statement in his *Treatise of Human Nature* (1739), it has been associated with the name of the Scots philosopher David Hume. It can be broadly stated in this way: when one infers from the observed O to the unobserved U, O and U are always logically distinct, at least in the sense that one can conceive of O holding, yet U not. So there is no logical necessity for U to follow from O. What then could form the grounds for asserting U, given O? For an empiricist (such as Hume) there was nothing that one could observe which would fit the bill – possible stopgaps such as natural necessity or causal powers were simply metaphysical fictions. There was, in short, merely a succession of events, and nothing we can observe guarantees that the unobserved will continue the pattern of the observed. Furthermore, any attempt to justify induction by a deductive argument would be inappropriate, for induction is essentially ampliative, in that the conclusion goes beyond what is contained in the premises, whereas deductive inferences are always conservative. Conversely, an inductive justification of induction, on the grounds that it has worked well so far, would appear to be circular.

The philosophical responses to this problem can be of two kinds. One response is to acknowledge that inductive inferences are unjustifiable, and that consequently they should play no role

in a rational enterprise such as science. Thus many authors have placed great emphasis on eliminative methods, whereby various potential explanations of the inductive evidence are eliminated as impossible or highly improbable, using primarily deductive methods. The best-known modern advocate of this view is Karl Popper, whose *Logic of Scientific Discovery* (1959) is, in part, a sustained defence of a purely deductive scientific methodology. Mill's famous methods of experimental inquiry (1843, Book III) are eliminative, as is part of Keynes's (1921) theory of induction, and a large portion of Bacon's approach. It is also possible to view in this way the objectivist statistical methodology of hypothesis testing, where the emphasis is on the rejection of statistical hypotheses. There is serious doubt with all of these approaches, however, as to whether they can function properly without tacitly employing inductive methods at some stage.

The second, and more common response is to provide some reasons why inductive inferences are indeed rationally justifiable. The 'missing premise' approach, for example, suggests that we view inductive arguments as incomplete deductive arguments, or enthymemes. By adding some extra assumption, usually a variant of a uniformity of nature principle, one can convert inductive arguments into deductive. Holders of this view have often felt compelled to adopt such a uniformity of nature principle as an a priori truth, one without which science would be impossible. The problems with this approach are many, primarily: what exact form should the uniformity of nature principle take, and how is it to be justified? 'The future resembles the past' is too vague, and almost certainly false, for dissimilarities are at least as common as similarities. 'Every event falls under a law of nature' may be true, but which law for which event? Mill tried to solve the problem by claiming that all inductions were inferences from particulars to particulars, although also asserting that a general uniformity of nature principle could be established inductively using the success of many more specific inductive generalizations.

The pragmatic approach to induction, credited to Hans Reichenbach (1949, pp. 469–82), argues that while induction cannot be guaranteed to work, if any method succeeds then induction will do just as well. Hence one may as well employ what Reichenbach called the 'straight rule', – infer that the relative frequency of observed positive instances of an effect will continue in the future. There is, however, an infinite number of alternative rules that are consistent with Reichenbach's procedure, and hence the vagueness problem is still with him.

Much philosophical work was done in the middle part of this century to construct systems of inductive logic using a logical probability function i.e. a numerical function which attributes a degree of inductive confirmation to an hypothesis, given certain evidence statements. This work, the most developed of which was carried out by Rudolf Carnap (1950), is generally regarded as having failed to achieve its aims. It did, however, produce a number of useful insights into the nature of inductive inferences, among which was the principle of total evidence, which asserts that in applications of inductive logic, no relevant evidence should be omitted.

Contributions of Economists

Among those who have made important contributions to both economics and the study of induction, we may count primarily J.S. Mill (1843), W.S. Jevons (1874), J.M. Keynes (1921), and R.F. Harrod (1956). Economists who have also written explicitly on induction include A.A. Cournot, F.Y. Edgeworth, F.P. Ramsey, John Hicks, Herbert Simon, and F.A. Hayek. (It is worth mentioning that Hume himself made a seminal contribution to economics with his theory of gold-flow equilibrium and defence of free trade.) Mill's views have been described earlier. Jevons's principal work is *The Principles of Science* (1874), within which the use of the hypotheticodeductive method is heavily stressed, as well as the allocation of subjective probabilities to those hypotheses by means of inverse probability methods. Jevons's inductive views are now for the most part regarded as combining exceptional insight with generally fallacious reasoning.

Keynes's only philosophical book, *A Treatise on Probability* (1921), is, like most of his work, of great originality. Here one can find one of the first systematic expositions of logical probability. Keynes is also perhaps the first to have insisted that logical probabilities are relative to evidence and cannot be separated from such. Hence there is no rule of detachment for probabilistic inductive logic, in the sense that evidence premises cannot be detached from the inductively supported conclusion, as is possible in deductive logic. This work on inductive logic was an important precursor of Rudolf Carnap's contributions in this area. Keynes also introduced the Principle of Limited Independent Variety, which essentially asserts that all inductive inferences concern objects with a finite number of independent properties, or, that there cannot be an infinite plurality of causes for an effect. This principle was necessary in order to attribute finite prior probabilities to the hypotheses under consideration. Harrod's (1956) theory cannot be swiftly stated: suffice it to say that he argues for the intrinsic acceptability of certain inductive arguments based on probability without supplementation by additional assumptions. His work has not attracted wide support.

Prospects

Is there a solution to Hume's problem? A characteristic of both kinds of approach discussed above has been their tendency towards an increased level of abstraction, symbolized by increasingly powerful mathematical and logical techniques. Useful as these techniques are, inductive inferences can rarely be made confidently without careful attention to causal relationships. Hume's problem itself arose directly from his argument that there is nothing more to causal connections than the regular succession of temporally ordered contiguous events. Mill gave careful attention to the causal foundations of inductions, but many empiricists are uneasy with causal talk, and the 20th century has largely eschewed causes in favour of mathematics. Because induction, causality and probability are so intimately connected, one may be able to rectify this neglect by making use of a specifically causal concept of probability (e.g. Humphreys 1985). That is, rather than construing probabilities as logical relations, subjective degrees of belief, or relative frequencies, one may take them to be propensities, i.e. probabilistic dispositions whose concrete structural basis is the economic system under investigation. Indeed, much of the work by Marschak, Hurwicz and by Simon (1977) on identifiability of structural parameters within causally isolated systems lends itself to this kind of approach. Those theories are ultimately reliant upon an understanding of causation which comes from experimental interventions, and since we are undoubtedly acquainted with primitive causal relations in that way, using such relations to justify others will not result in circularity. By localizing such inferences, there need be no vagueness about the inductive claims made. This approach does suffer from the extreme difficulty of identifying causal relationships within complex economic systems, and this difficulty is, of course, why one often must replace the experimental controls of simpler physical sciences by statistical surrogates for economic purposes. Complete certainty about inductive inferences is impossible, but the clear and discoverable differences between stable and unstable systems, equilibrium and disequilibrium, and isolated and non-isolated systems lie at the heart of the difference between secure and insecure inductive inferences from the past to the future, and a judicious mixture of statistical techniques with causal models seems to offer a promising alternative to the acausal inductive heritage of Hume.

A comprehensive bibliography up to 1921 may be found in Keynes (1921). More recent work is cited in Swinburne (1974). The best survey is still Kneale (1949) and an elementary source is Skyrms (1986).

See Also

► Analogy and Metaphor
► Hume, David (1711–1776)

Bibliography

Bacon, F. 1620. *Novum Organum*. Reprinted as *The new organon*, ed. F.H. Anderson. Indianapolis: Bobbs-Merrill, 1960.
Barnett, V. 1982. *Comparative statistical inference*. 2nd ed. Chichester: John Wiley.
Carnap, R. 1950. *Logical foundations of probability*. Chicago: University of Chicago Press.
Harrod, R. 1956. *Foundations of inductive logic*. New York: Harcourt, Brace.
Hume, D. 1888. In *1739. A treatise of human nature*, ed. L.A. Selby-Bigge. Oxford: Oxford University Press.
Humphreys, P. 1985. Why propensities cannot be probabilities. *Philosophical Review* 94: 557–570.
Jevons, W.S. 1874. *The principles of science*. London: Macmillan.
Keynes, J.M. 1921. *A treatise on probability*. London: Macmillan.
Kneale, W. 1949. *Probability and induction*. Oxford: Clarendon Press.
Mill, J.S. 1843. *A system of logic*. London: J.W. Parker.
Popper, K. 1959. *The logic of scientific discovery*. London: Hutchinson.
Reichenbach, H. 1949. *The theory of probability*. Berkeley: University of California Press.
Simon, H. 1977. Causal ordering and identifiability. In *Models of discovery*, ed. H. Simon. Dordrecht: D. Reidel.
Skyrms, B. 1986. *Choice and chance*. 3rd ed. Belmont: Wadsworth.
Swinburne, R. 1974. *The justification of induction*. Oxford: Oxford University Press.

Industrial Organization

Richard Schmalensee

Based on the activities of those who consider themselves in the field, industrial organization (or industrial economics) today may be broadly defined as the field of economics concerned with markets that cannot easily be analysed using the standard textbook competitive model. In such markets the positive and normative implications of models of imperfect competition are generally of interest, as are the design and effects of government antitrust and regulatory policies aimed at improving market performance. Because there are many models of imperfect competition, and because general policies must be applied to particular cases, much of the research in industrial organization has been and continues to be empirical.

Historically, industrial organization emerged as a distinct field after the rise of the modern manufacturing enterprise around the turn of the century (compare Chandler 1977 and Hay and Morris 1979, ch. 1). Early writers largely equated 'industrial' with 'manufacturing' and focused on markets for manufactured products. Students of industrial organization today do not limit themselves exclusively to the manufacturing sector, but, in part because of the availability of data, departures from that sector are selective.

Thus securities markets, which seem to approximate perfect competition well, are not studied in industrial organization, but competition among financial institutions and regulation of their behaviour have been investigated. Studies of transportation and traditional public utilities are common, in part because of the important role played by government policy in these sectors.

Industrial organization has also retained a strong focus on the firm as an object of study. In microeconomic theory, the firm is a given cost or production function assumed to be operated to maximize profits; in industrial organization the structure and behaviour of firms are objects of study. In contrast, relatively little attention is devoted to household behaviour.

The national markets for manufactured goods that were created early in this century have two important and apparently novel characteristics, stressed in Chamberlin's (1933) seminal and controversial analysis. First, products in many manufacturing markets are *differentiated*; that is, buyers do not view them as perfect substitutes. In such markets, non-price competition, involving product design, advertising, and other selling expenses, is often important. The sources and consequences of product differentiation and non-price competition have been intensively studied.

Second, some industrial markets came to be dominated by a relatively small number of firms. A good deal of work in industrial organization has attempted to explain differences in the

'organization' of markets, focusing on *seller concentration*, the extent to which sales are concentrated in the hands of a small number of firms. The consequences of seller concentration have also been intensively studied, and the analysis of oligopoly behaviour has accordingly played a central role in industrial organization.

The main objective of the field has been to develop tools to analyse market processes and their consequences for economic performance. Since Bain (1959), it has been customary to work with the concepts of structure, conduct, and performance. Market *structure* refers to a set of variables that are relatively stable over time, observable (at least in principle), and that are important determinants of buyer or seller behaviour. All scholars in the field agree implicitly or explicitly that there exists such a set of variables; otherwise market behaviour is in principle unpredictable.

Intrinsic market structure variables (termed *basic conditions* by Scherer 1980, ch. 1) are essentially completely determined by the nature of the product and the available technology: all modern steel industries are capital-intensive, for instance. Other elements of market structure are *derived* in that they may reflect government policy, corporate strategies, or accidents of history: the concentration that was created by the US Steel merger in the USA in 1901 is an obvious example (see Chandler 1977, ch. 11 and Stigler 1968, ch. 9). Intrinsic structural variables also affect derived variables to some extent: even if the US Steel merger was not inevitable, it is difficult to imagine an atomistic steel industry. The strength of these effects is perhaps inevitably controversial, since the stronger they are, the less scope there is for governments to enhance efficiency by changing market structures.

In any complete market model, market structure determines the *conduct* of buyers and sellers. Compare, for instance, structure and conduct (rules for output choice) in pure monopoly and in perfect competition. An important objective of industrial organization is to describe and predict the conduct of actual industries in terms of a continuum joining these two polar cases: as one moves toward the competitive end, the intensity of *rivalry* increases and profits fall accordingly. But a single dimension cannot describe conduct fully; market behaviour typically involves choosing which products to produce, the corresponding vector of prices or outputs, distribution and advertising strategies, and levels and directions of research and development activity.

Market *performance* is assessed by comparing the results of market behaviour in efficiency terms to first-best ideals or feasible alternatives. One might compare prices with marginal costs, for instance, or the array of products produced with some ideal array. Performance is determined by all aspects of conduct, along with the intrinsic elements of market structure.

To this relatively static framework, one must add dynamic effects of buyer and seller behaviour on market structure. Intrinsic structural variables can be changed by innovation, for instance, and seller concentration can be changed by mergers. Established sellers may be able to take actions to inhibit the entry of new rivals.

Much early work in industrial organization eschewed formal theory, in part because there did not exist an adequate general theory of behaviour in oligopolistic markets. Many scholars concentrated on induction from case studies of particular markets. At the same time, others sought to reinterpret and extend the standard competitive and monopoly models to enhance their explanatory power. Beginning in the 1950s, cross-section statistical work on samples of manufacturing industries became common. In the 1960s, particularly in the USA, industrial organization economists began to look beyond antitrust policy and to examine systematically the effects of government regulatory programmes.

Recently, formal models of imperfect competition have been studied intensively, and the tools of noncooperative, extensive form game theory have assumed central importance in this work (see Schmalensee 1982; Waterson 1984; Roberts 1985). Laboratory experiments are being performed more frequently, as are case studies relying heavily on formal models and econometric analysis of firm behaviour. Implications of imperfect competition for international trade and for national welfare in a world economy are receiving increased attention.

In the remainder of this essay I discuss briefly some of the questions that have been studied intensively by industrial organization economists. For more detailed treatments of many of these issues, see Bain ([1959], 1968), Hay and Morris (1979), Scherer (1980), Schmalensee and Willig (1988), Stigler (1968), and Waterson (1984). Space constraints preclude an explicit discussion of antitrust or regulatory policy.

Are firms managed so as to maximize profit (or, more generally, the wealth of their owners), as microeconomic theory assumes? This question is most important in imperfectly competitive markets, since non-maximizing firms cannot survive in the long run under perfect competition. Two alternatives to profit-maximization have been advanced. Neither has yet proved to be more generally useful, though both yield valuable insights in some situations.

First, some scholars have argued that firms' problems are so complex and their information so imperfect that maximization is effectively impossible. They stress the importance of routines, rules of thumb, experimentation, and learning in actual business behaviour.

Second, others note that the many shareholders who nominally own large corporations cannot effectively review managers' decision-making. These scholars model managers as pursuing a variety of their own objectives (such as firm size or growth) subject to constraints (often relating to profitability) imposed by owners. Recent work using agency theory to model the manager–owner relation shows considerable promise here.

In an ideal world, shareholders would always use the *market for corporate control* to replace managers who did not effectively pursue the owners' interests. And shareholders in fact often force mergers and sometimes elect boards of directors opposed by incumbent managers. But ours is not an ideal world, and the importance of frictions and imperfections in the market for corporate control is widely debated.

What determines the boundaries between firms and markets? A number of authors have focused on the implications of cost minimization for firm structure. Oliver Williamson (1975) has stressed the difficulty of writing long-term contracts that allow for all possible contingencies. If efficient production requires making investments that cannot be easily shifted to alternative uses, this difficulty may make it more efficient to integrate related activities within a single firm, rather than to attempt to coordinate them by contract and risk the effects of contractual breakdown.

Stigler (1968, ch. 12) stressed the importance of *economies of scale*. He argued that, as markets grew, specialized firms would arise to perform functions in which economies of scale were important. More recently, Baumol et al. (1982) have argued that multiproduct firms may arise to take advantage of *economies of scope*, which lower cost when the production of multiple products is carefully coordinated. Scope economies arise when assets can be readily shared among processes producing several outputs.

Another even more diverse body of theoretical literature argues that imperfections in competition may produce other incentives for firms to expand their activities. Relatively few empirical tests of any of these models have been performed, however.

What is a market? In microeconomic theory, a market is the locus of trades in a single, perfectly homogeneous product. This definition would make almost all real firms monopolists. In practice, markets must be defined by aggregating products that are relatively close substitutes in demand or supply. For most purposes, particularly in the context of antitrust policy, it is useful to define a market as the smallest aggregate that could profitably be monopolized. It is rarely easy to implement this definition in a fully satisfactory way, however, and government datacollection agencies rarely try. This poses real problems for empirical work.

What are the key elements of market structure? In the seminal work on this point, Bain (1959) argued that there were four such elements, all of which he seemed to treat as derived: the extent of seller concentration, the extent of buyer concentration, the importance of product differentiation, and the conditions of entry. Seller concentration was held to facilitate noncompetitive behaviour; buyer concentration was held to make such behaviour harder to sustain. Bain argued that product differentiation insulated sellers from each others' actions and changed the focus of rivalry from

price to non-price competition. He also argued that the easier it was for new competitors to enter an industry, the more difficult it would be for established firms to maintain prices above costs and earn supra-normal profits. (Baumol et al. (1982) have coined the term *contestable* to describe markets in which entry is so easy that potential competition alone suffices to eliminate excess profits.)

Bain (1956) went on to identify four sources of *barriers to entry*, four reasons why established firms might be able to earn excess profits without facing the threat of entry. (See Stigler 1968, ch. 6 for an important alternative definition of this term.) First, substantial economies of scale might make potential entrants reluctant to enter at efficient scale for fear of depressing prices below costs. Second, established firms might have cost advantages over potential entrants, perhaps because of proprietary production processes. Third, established firms might have demand-side or product differentiation advantages over potential entrants, perhaps because of patented products or buyers' reluctance to switch brands. Finally, Bain felt it was possible that imperfections in capital markets would inhibit entry when large initial investments were required. This last possibility remains controversial.

While this framework remains influential, it is increasingly under attack. Bain and many of his followers assigned what now seems to be excessive importance to seller concentration. Bain's framework neglects firm structure, but firms that operate in many markets may behave differently from single-market enterprises, since actions taken in one market may affect costs or strategic opportunities in others. Caves and Porter (1977) have argued that the notion of entry barriers must be generalized to include *mobility barriers*, which impede entry into *strategic groups* of sellers with similar capabilities and strategic objectives. They and others contend that the structure of strategic groups within industries can materially affect conduct.

Bain's framework now seems to many scholars to omit a number of critical structural variables. Recent work attaches particular importance to cost conditions and information. Baumol et al. (1982) have stressed the impact of *sunk costs*, costs required to enter a market that cannot be recovered if the market is later abandoned. The more important sunk costs are, the greater the risk of entry, and the more important scale economies are as a barrier to entry, all else being equal.

Stigler (1968, ch. 5) pointed out that sellers are more likely to be able to sustain non-competitive behaviour the better their information on each others' actions. Recent game-theoretic work has expanded on this insight and stressed the importance of information about rivals' capabilities and objectives as well. Buyers' information about prices and qualities may also play a central role in determining marketing and distribution arrangements and the form and intensity of rivalrous behaviour. If buyers must spend time to learn the prices of competing sellers, for instance, each seller has some monopoly power even if there are many firms marketing identical products.

Bain ([1959] 1968, p. 9) argued that one ought to restrict attention to a small number of structural characteristics because 'meaningful intermarket comparisons and meaningful generalizations about the influence of structure on behaviour are effectively forestalled if the content of "structure" is made so comprehensive that no two markets could be viewed as structurally alike'. But many scholars now feel that simple generalizations of the type that Bain sought may not have much predictive power.

How are the derived elements of market structure determined? Most work has focused on the determinants of seller concentration, with special attention given to the hypothesis that concentration is determined by economies of scale. This hypothesis is supported by the observation that the same industries tend to be concentrated in all developed economies, despite different histories and government policies.

Scale economies in manufacturing at the plant and firm levels have been measured by statistical methods, by interview studies (the 'engineering' approach), and by comparing the sizes of units that prosper and decline (the 'supervisorship' approach'. These studies have been criticized because of the inherent difficulty of measuring non-production scale economies (those in

marketing and distribution, for instance) that occur at the firm level. In general this work suggests that concentration in most US manufacturing industries is higher than required for the exploitation of economies of scale in production. This is consistent with the observation that among large industrialized economies, absolute levels of concentration in particular industries are not sensitive to differences in the size of the national market.

A number of other potential sources of concentration have been identified. Spence (1981) has shown that *economies of learning*, which cause unit cost to decline with cumulative production, can mandate high concentration even when scale economies are absent. Demsetz (1973) has argued that persistent efficiency differences, along with the tendency for efficient firms to expand at the expense of their rivals, are an important source of concentration. Many others have studied the impacts of random variations in firm growth rates and of mergers on concentration. Mergers are an important source of concentration in some countries but not in others; the empirical importance of the other factors remains controversial.

What determines the intensity of rivalry in oligopolies? We still have no fully satisfactory, general model of oligopoly, but theoretical work has yielded a number of valuable insights.

The basic problem faced by any set of sellers is that posed by the classic prisoners' dilemma game. In a static setting, all sellers do well if prices are kept high. If all other sellers set high prices, however, any single firm can usually increase its profits by charging a lower price (or producing more than its assigned quota). If all behave selfishly in this fashion, all will charge low prices and receive low profits. That is, *non-cooperative*, selfish behaviour in this setting tends to produce competitive, low-profit outcomes.

Interest thus attaches to the possibility of *cooperative* or *collusive* behaviour that can produce monopolistic performance, with high prices and high profits. In principle, collusive behaviour can be overt, with firms explicitly agreeing on strategies, or tacit, with firms reaching an unspoken understanding about acceptable policies. It is more difficult to reach agreement tacitly than overtly; agreement is also more difficult the more firms there are and the greater the differences among them. Collusion can either be total, covering all decisions, or partial, covering only some variables under firms' control. It may be easier to collude on price than on advertising, for instance.

Collusive agreements are inherently unstable, since individual sellers can usually increase their profits, as in the prisoners' dilemma game, by departing unilaterally from the agreement. Stability requires the ability to detect cheating and to make a *credible threat* (one that it would actually be rational to carry out) to impose a sufficiently severe penalty to render cheating unprofitable. The game-theoretic notion of *perfect* (Nash) *equilibrium,* in which players noncooperatively pursue their own interests but noncredible threats are ruled out, has been used heavily in recent work on oligopoly theory and entry deterrence (discussed below).

Stigler (1968, ch. 5) argued that cheating can be more reliably detected in concentrated markets. A number of authors have recently built on his work and devised multi-period gametheoretic models in which firms announce credible threats that make cheating irrational, even when cheating can only be imperfectly detected. In these models threats are sometimes carried out (price wars occur) even though nobody ever cheats.

A number of econometric time-series studies of individual oligopolistic markets have been undertaken in recent years. These often employ the non-game-theoretic formalism of *conjectural variations*. In a market in which products are undifferentiated and firms set outputs, a firm's conjectural variation is its expectation of the derivative of all other firms' output with respect to its own. Estimates of the conjectural variations consistent with observed market outcomes provide a summary description of the intensity or rivalry. If all conjectural variations equal minus one, behaviour is perfectly competitive; larger values imply departures from the competitive ideal. Data limitations make it hard to apply this approach to many industries.

Can the conduct of established sellers discourage the entry of new rivals? In the classical limit-

pricing model of Bain (1956), an established monopoly in an industry with significant scale economies could discourage entry by raising output above the monopoly level and threatening not to reduce production if entry occurred. The incumbent would optimally select its output so that entry at efficient scale would raise total output so much as to depress price (just) below cost.

Unfortunately, the threat in this model is not credible (i.e., the no-entry equilibrium is not perfect). If entry did occur, the incumbent could generally increase its own profits by reducing its output, and a potential entrant has no reason to believe that the incumbent would forego such an opportunity. Recent work (see, especially, Roberts 1987) imposes the requirement of credibility. There are three strands to this literature.

First, when information is imperfect, incumbent firms may attempt to use pre-entry price to deceive potential entrants. If potential entrants don't know the incumbent's costs, for instance, the incumbent may lower its pre-entry price below the monopoly level in order to persuade potential rivals that its costs are too low to permit viable entry. This resembles classic limitpricing, but it turns out on average not to deter sophisticated entrants, who understand the incumbent's incentive to attempt deception.

Second, an incumbent may be able to make credible threats by making *commitments* in advance of entry. That is, it may be able to take actions before entry that alter its post-entry incentives in a way that makes a hostile response to entry more attractive. Spence (1977), who began this line of work, considered investment in production capacity as a vehicle for commitment.

Third, if entrants are uncertain about an incumbent's objectives, it may be rational for the incumbent to take predatory actions designed to eliminate entrants when they appear. Such a policy may give it a *reputation* for aggressive (or irrational) behaviour, which may serve to deter even sophisticated potential entrants.

Do statistical analyses of data from multiple industries shed light on the validity of the hypotheses discussed above? Many cross-section studies have been performed by students of industrial organization, but the interpretation of many of their statistical findings is controversial, and the intertemporal stability of some key relationships has recently been questioned.

All cross-section studies employ accounting data, and most focus on determinants of profitability. But accounting data do not provide exact measures of real, economic profitability, in part because of differences in riskiness, the way in which long-lived investments are depreciated, and the possible ability of labour unions to capture rents generated by collusive behaviour. It has proven difficult to obtain good measures of other theoretical constructs as well, particularly product differentiation and barriers to entry. The growing importance of firms operating in several markets poses yet another measurement problem.

Many cross-section studies find a weak but statistically significant positive relation between seller concentration and industry profitability. Until recently this was generally interpreted as supporting the Bainian hypothesis that concentration facilitates collusion. But Demsetz (1973) has offered an alternative explanation: in a world without collusion, substantial efficiency differences among rival sellers are likely to produce both concentration, as noted above, and high industry-level profits, because efficient firms earn rents. Where efficiency differences are unimportant, one would expect both concentration and profits to be low. This hypothesis is consistent with the strong positive correlation between market share and profitability in some industries, but not many industries follow this pattern. It has proven difficult to discriminate between these two hypotheses empirically.

Similarly, numerous studies have found a strong positive correlation between advertising/sales ratios and profitability. This has often been taken to support Bain's (1956, 1959) hypotheses about the effects of product differentiation and product differentiation advantages of established firms. But advertising is logically only one input affecting those structural variables, and it is an endogeneous variable, determined by profit-seeking sellers. Unfortunately, it has been difficult to specify good simultaneous equations models in this area. Finally, the effects of advertising on demand probably persist over time, so that

advertising should be treated as an investment, not a current expense. If advertising's effects are assumed generally to decay slowly enough, the correlation between advertising intensity and (corrected) profitability measures disappears, but the appropriate decay rate assumption remains controversial.

Because of these and other problems of measurement and interpretation, inter-industry empirical work seem to have lost the central place it formerly held in industrial organization. Still, such work remains an important source of the general stylized facts needed to guide the construction of useful theoretical tools.

What sorts of price structures are imposed by firms with market power? What are the welfare implications of price discrimination? If a seller has some control over its price (i.e., it is not a perfect competitor), can identify (even imperfectly) customers with different demand characteristics, and can prevent (or at least inhibit) trade among its customers, it will generally pay it to practise price discrimination. That is, it will adopt a price policy in which different customers pay different marginal or average prices depending on their demand characteristics. The ability to earn excess profits is not required; price discrimination can persist in a free-entry equilibrium of the Chamberlin (1933) type.

In practice, sellers employ many devices for identifying customers of different types. Bulk discounts, in which the average price paid falls with volume, provide one method. (This is a special case of *non-linear pricing*, in which the buyer's bill is a nonlinear function of the quantity he purchases.) Delivered pricing, in which a buyer's price depends on his location, provides another. Discrimination may also be effected by bundling or tying arrangements, which require buyers to purchase related products from a single seller. And there are a host of market-specific devices: US airlines, for instance, charge a much lower fare for trips that involve spending a Saturday away from home, thus generally charging lower prices to tourists than to business travellers.

The large theoretical literature on the consequences of such practices contains few sharp results. Prohibiting price discrimination in most cases produces both gainers and losers; the net welfare effect is usually ambiguous.

How are product quality and variety determined in imperfect markets? Are the outcomes likely to be optimal in any sense? If 'quality' is simply inserted as an additional variable in a standard monopoly model, one can show that 'quality' may be either too high or too low in equilibrium, depending on the details of the demand function. If there are many sellers and buyers rely on firms' reputations for quality in making decisions, firms producing high quality (and thus high cost) products must be able to charge prices above marginal cost in equilibrium. If not, they would have an incentive to lower quality and exploit their reputations until buyers caught on.

Models in which variety is determined usually assume economies of scale in the form of brand-specific fixed costs; otherwise it would generally be socially efficient and privately optimal to produce all possible brands. This assumption rules out purely competitive equilibria and forces second-best welfare comparisons. It also provides a rationale for intra-industry international trade: such trade expands the market and makes greater variety economically feasible.

In some models of variety determination, the demand side of the market is a single representative consumer who desires variety. In others, consumers have different ideal brands, and each desires, all else equal, to consume the brand that is 'closest' to his ideal in the space of all possible brands. In a third class of models, consumers agree on the ranking of all possible brands but differ in their willingness to pay for quality. Market equilibria in all these models generally involve a non-optimal set of brands, but the nature of deviations from optimality depends on the details of the model. Chamberlin's (1933) view that monopolistic competition implies excessive variety is not generally valid.

Are market-determined levels of advertising excessive? Do they increase barriers to entry? Neither of these traditional questions has yet been answered, and neither may in fact have a general answer.

In order to assess the optimality of any level of advertising, one must make some assumption

about how advertising affects consumer behaviour. One extreme assumption is that advertising simply provides consumers with information; the other is that it simply changes their tastes. Under the first assumption, market-determined advertising levels are optimal only under very special conditions; under the second assumption the optimal level of advertising depends on what tastes are used as a yardstick. In fact, neither extreme assumption is likely to be generally correct.

The theoretical effect of advertising on conditions of entry depends, again, on the way advertising affects consumers, and this is likely to differ among markets. In some markets restrictions on advertising are observed to increase prices; in others, heavily advertised brands sell at substantially higher prices than apparently physically identical brands that are not advertised. Advertising may be less important in markets in which retailers are an important source of information. Product differentiation advantages of established brands may depend more on satisfied buyers' rational reluctance to experiment with new brands than on the effects of advertising.

Are large firms in concentrated markets the major sources of technical progress, as Schumpeter (1942) *argued*? It is difficult to measure any firm's contribution to technical progress; counts of patents or significant innovations are frequently used but obviously imperfect indicators. Most studies have found that, in most industries, large firms are not disproportionate sources of innovations, especially not significant innovations, but there are exceptions. It is more difficult to assess the impact of market structure on technical progress, since one must control for differences in the opportunities for innovation across markets. The available evidence provides at most weak support for Schumpeter's view of the effects of concentration.

On the theoretical side, a number of authors have recently modelled research and development rivalry in game-theoretic terms. In many of these models, firms spend money (perhaps over time) to increase their chances of winning a single prize, usually interpreted as a patent. Under some conditions, it may be rational for an incumbent monopolist to outspend potential entrants in order to prevent their entry. In general, theoretical work indicates that market-determined levels of research and development spending may be excessive or inadequate. This work also makes clear that concentration and other structural variables are in the long run determined by the intrinsic opportunities for innovation. Concentration and innovative activity are thus both endogeneous variables.

How important are departures from competitive performance? Early studies of this question, which associated differences in profit rates with departures from the competitive ideal, concluded that monopoly power imposed relatively small costs on society. It is now clear that a proper general equilibrium analysis of this issue may imply much larger or even smaller effects, depending on the values of unknown parameters.

Some authors have argued that differences in observed profit rates understate the actual effects of monopoly power because monopoly profits are to some extent dissipated in actions taken to achieve or protect monopoly positions, captured by labour unions, or simply foregone by lazy or inept managers not subject to market discipline. On the other hand, if Schumpeter (1942) was right, short-run measures of the cost of monopoly omit important long-run benefits. Like so much else in this field, the actual importance of departures from pure competition in modern economies remains controversial.

See Also

- ▶ Advertising
- ▶ Market Structure
- ▶ Selling Costs

Bibliography

Bain, J.S. 1956. *Barriers to new competition*. Cambridge, MA: Harvard University Press.

Bain, J.S. 1959. *Industrial organization*, 2nd ed. New York: John Wiley. 1968.

Baumol, W.J., J.C. Panzar, and R.D. Willig. 1982. *Contestable markets and the theory of industrial structure*. New York: Harcourt, Brace, Jovanovich.

Caves, R.E., and M.E. Porter. 1977. From entry barriers to mobility barriers. *Quarterly Journal of Economics* 91(2): 241–261.
Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.
Chandler, A.D. 1977. *The visible hand: The managerial revolution in American business*. Cambridge, MA: Harvard University Press.
Demsetz, H. 1973. Industry structure, market rivalry, and public policy. *Journal of Law and Economics* 16(1): 1–9.
Hay, D.A., and D.J. Morris. 1979. *Industrial economics: Theory and evidence*. Oxford: Oxford University Press.
Roberts, D.J. 1987. Battles for market share: Incomplete information, aggressive strategic pricing, and competitive dynamics. In *Advances in economic theory II*, ed. T. Bewley. Cambridge: Cambridge University Press.
Scherer, F.M. 1980. *Industrial market structure and economic performance*, 2nd ed. Chicago: Rand-McNally; 1st edn, 1970.
Schmalensee, R. 1982. The new industrial organization and the economic analysis of modern markets. In *Advances in economic theory*, ed. W. Hildenbrand. Cambridge: Cambridge University Press.
Schmalensee, R., and R.D. Willig (eds.). 1988. *Handbook of industrial organization*. Amsterdam: North-Holland.
Schumpeter, J.A. 1942. *Capitalism, socialism, and democracy*. New York: Harper.
Spence, A.M. 1977. Entry, capacity, investment and oligopolistic pricing. *Bell Journal of Economics* 8(2): 533–544.
Spence, A.M. 1981. The learning curve and competition. *Bell Journal of Economics* 12(1): 49–70.
Stigler, G.J. 1968. *The organization of industry*. Homewood: Irwin.
Waterson, M. 1984. *Economic theory of the industry*. Cambridge: Cambridge University Press.
Williamson, O.E. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Industrial Relations

Thomas A. Kochan

Abstract

Industrial relations is an interdisciplinary field of study that encompasses all aspects of work and employment relations. Originating in institutional economics and Fabian socialism, it has evolved to address employment problems and issues ranging from wage determination and collective bargaining to human resource management and labour market dynamics and policies. Globalization has increased the importance of international and comparative analysis of employment practices and outcomes. Shifts in employment from manufacturing to services has rendered the term 'industrial relations' obsolete and led scholars to use the term 'work and employment relations' to describe their work and this field of study.

Keywords

American Economic Association; Arbitration; Class conflict; Collective bargaining; Commons, J. R.; Corporate governance; Corporations; Ely, R. T.; Fabian socialism; Globalization; Great Depression; Health insurance; Human capital; Human relations; Human resource management; Incentive compensation; Industrial psychology; Industrial relations; Innovation; Institutional economics; Internal labour markets; Labor and Employment Relations Association (USA); Labour economics; Layoff; Marx, K. H.; Minimum wages; Networks; New Deal; Pensions; Personnel economics; Power; Scientific management; Social insurance; Social networks in labour markets; Strikes; Team production theory of the firm; Technical change; Trade unions; Training; Unemployment insurance; Wage determination; Webb, S. and B.; Women's work and wages; Workers' compensation

JEL Classifications

J5

Industrial relations is an interdisciplinary field devoted to the study of all aspects of work and employment relations. It emerged historically out of the works of the Fabian Socialists Sydney and Beatrice Webb (1894, 1897) in Great Britain and institutional economists such as John R. Commons (1909, 1934) in the United States. Both sets of scholars and their students were searching for ways to understand *and* influence

employment relations in ways that distinguished their normative, theoretical and methodological approaches from Marx (1849) on the one hand and classical or neoclassical economics (Marshall 1920) on the other.

Over the years the field has evolved and broadened considerably to incorporate concepts and methods from other social sciences such as psychology, sociology, and political science and from disciplines outside the social sciences such as history and law. In recent years the term 'industrial relations' has become somewhat dated, given the growth of the service sector and the decline of traditional manufacturing industries, leading a number of research units working in this scholarly tradition to redefine the field as the study of 'work and employment relations'. But the underlying normative, theoretical and methodological features of the field carry on the distinctive features of industrial relations.

Origins and Initial Intellectual Debates

Karl Marx provided the intellectual rationale and stimulus to the field of industrial relations. His most enduring contribution was to assert that labour was more than just a commodity or factor of production subject to deterministic laws of supply and demand. Instead, the free will and power that reside in human beings make labour more than an inanimate object. This basic insight serves as an enduring normative premise in industrial relations and motivates much of the work in the field to this day. That is, while affected by market forces similar to other factors, labour deserves and requires special treatment in theory and public policy because workers can take individual or collective actions to influence market outcomes, and work and employment relationships affect important human values and have important social as well as economic consequences. For these reasons, industrial relations research, public policies and practices need to be as concerned about equity as efficiency at work (Barbash 1984; Meltz 1989). Moreover, freedom of association at work is recognized as a fundamental human right in democratic societies and, therefore, the ability of workers to have a voice in determining their employment conditions serves as an equally important industrial relations outcome (Budd 2004).

While Marx provided the starting point for the field of industrial relations, much of the scholarship in the field has taken issue with other aspects of Marxian analysis. This is especially true of the Marxian view of the source of labour conflict in employment relations. Marx saw conflict at work as inevitable and all-encompassing, arising out of class differences rooted in the capitalist system of production. Conflict could be eliminated only by the revolutionary overthrow of that system. This became a major point of differentiation between Marxist and labour process schools of industrial relations on the one hand (Hyman 1975) and on the other hand the more mainstream pluralist model which has grown to dominate European and Anglo-Saxon research traditions (Clegg 1970; Fox 1971; Kochan 1980).

Sidney and Beatrice Webb (the Webbs) were among the first to challenge Marx with their model of Fabian socialism. They shared with Marx a concern for the plight of the growing working class. Beatrice Webb was a student of the factory conditions prevailing in 19th century as Britain ushered in the first Industrial Revolution. Her empirical observations of factory conditions convinced her that the average worker suffered from an inherent imbalance of power in dealing with his or her employer. Trade unions were therefore needed to provide increased social support and bargaining power. Over time, however, unions were expected to evolve into institutions that promoted orderly government regulation that worked to the common benefit of all workers and for the overall community. Thus through evolution, not the revolution predicted by Marx, societies would evolve to better balance the needs of workers, employers and the communities in which they were embedded.

At about the same time as the Webbs were doing their work in Britain a quiet revolt was taking place in the field of economics in the United States. Leading economists were frustrated with the highly deductive and mathematical features of late 19th-century economic research.

As a result, in 1886 Robert Ely at the University of Wisconsin led other colleagues to form a new American Economic Association in an effort to bring a more inductive, empirical, and institutional brand of economics to bear on the critical problems of the day. Labour economics, more than any other sub-field within the economics profession, took up the institutional approach. Led by Ely's protégé at Wisconsin, John R. Commons, a new field was born. It focused on the study of labour and working conditions using empirically based, inductive methodological methods and focused as much on the collective institutions and organizations of workers and employers governing work and employment relations as on the actions of individuals in response to market forces.

Like Marx, these institutionalists believed that labour was more than a commodity. But unlike Marx, Commons and those that followed in building the field of industrial relations saw the conflicts of interests between employers and employees as part of natural, legitimate and ongoing differences in economic interests, not as a function of the capitalist system. Employers have the responsibility of promoting efficient use of scarce resources, including labour. Employees have the right and need to pursue their self-interests, individually or collectively, to improve their security, wages, working conditions, and other features of their work lives they value. These conflicting interests are not, however, absolute. Employers and employees also have some common interests that tie them together in ongoing interdependent relationships. Both want to generate value from their relationships so that there is more value to share. Safety and security may be other shared values. Commitment to the mission of the organization and service to their clients, customers, patients and so forth may be other shared values and objectives. Thus employment relations involve an inevitable mix of separate, perhaps conflicting, and common or shared goals. The task of industrial relations theory, research, teaching and policy therefore focuses on both finding an equitable resolution of differences or conflicts and ways to support value, creating solutions where interests overlap or are held in common (Walton and McKersie 1965).

The early institutionalists were strong proponents of empirical research and active involvement in policymaking and institution building. They studied labour market dynamics and labour management relations through field work more than through deductive model building. Their collective body of research and personal involvement generated most of the ideas and policy proposals embedded in the labour legislation of the New Deal. Unemployment insurance, workers' compensation, child and women's labour protections, minimum wages and social security all were ideas developed and studied at state and local levels of the economy between 1900 and 1930. Commons is now widely recognized as the intellectual father both of the New Deal labour legislation and the study of industrial relations in America (Kaufman 1993).

Debates with Alternative Disciplines

Kuhn (1970) argues that a new paradigm for the study of a phenomenon must be judged ultimately by whether it is better able than its alternatives to solve problems. So it is appropriate to examine industrial relations against this criterion at critical stages in its development.

Scientific Management Scientific management and industrial engineering dominated the study and practice of management and the design of work systems in the United States in the first two decades of the 20th century. The objective was to use engineering principles to find the optimal, most efficient methods for carrying out tasks, organizing them into a clear hierarchy and controlling labour through appropriate economic incentives and supervision to conform to the specified work process. In following these scientific engineering principles one would eliminate any potential conflicts of interests at work (Taylor 1895). This view of work and employment relations saw no rationale for worker voice, representation, or policies that would balance power between workers and managers. Its primary

theoretical prediction was that efficient organization and supervision of work, when supported by the right individual incentive compensation system, would generate maximum efficiency. Because efficient work would be rewarded, it would in turn generate worker satisfaction. This virtuous cycle would keep conflict from emerging in employment relationships. Thus, scientific management theory and efforts to implement it in practice stood in sharp contrast to industrial relations theories and normative assumptions.

Industrial Psychology At the same time industrial psychology was emerging as a field of study that paralleled and complemented the engineering approach. The study of personnel management largely grew out of industrial psychology. In contrast with the institutional economists, individuals, not collective groups or organizations, were the central unit of analysis and the firm was viewed more as a closed system, on the assumption that management controlled workplace decisions. Institutionalists reflected their economics' training by treating work and organizational practices as influenced by both organizational and external market and technological forces.

Human Relations In the 1920s the field of human relations was born out of the Hawthorne experiments (social-psychological experiments conducted at the Hawthorne Works' plant of Western Electric) in group behaviour and gave rise to another competing paradigm for the study of work and employment relations. The human relations school focused on work groups as the key unit of analysis and the social dynamics that shaped worker attitudes and behaviour. Human relations theorists reversed the theoretical argument of scientific management by proposing that worker satisfaction drove efficiency at work rather than the other way around (Roethlisberger and Dickson 1939). This school of thought provided intellectual foundation for the emergence of welfare capitalism in the 1920s. Large firms sought to provide a set of benefits and positive working conditions in order to achieve efficiency and in the process of doing so eliminate the incentives of workers to join trade unions (Jacoby 1991).

These were the alternative paradigms competing for influence with industrial relations over the first 30 years of the 20th century. The Great Depression of the 1930s raised industrial relations ideas, policies and research to a more prominent and perhaps dominant place in the intellectual and policy debates about work and employment relations. With the rise in industrial conflict and massive unemployment came the recognition of the need to establish a floor on labour standards and a means for workers to bargain as relative equals with their employers to improve on these minimum conditions. Thus, it was the dramatic deterioration in economic conditions, the threats unregulated conflict posed to democracy and social stability, and the shift in the political environment that allowed the ideas and research evidence of the institutional economists to emerge as the intellectual basis for much of the New Deal legislation passed in the 1930s.

The New Deal Era, the Second World War and the War Labor Board

From 1932 to 1945 industrial relations scholars and practitioners had an unprecedented impact on national policy and private practices of employment relations in the United States. The War Labor Board (WLB) (1941–5) that was charged with controlling wages and mediating collective bargaining negotiations played a key role in legitimating and starting the long-term diffusion of many modern personnel and labour relations practices and benefits including grievance arbitration, cost of living wage increases, paid time off for holidays and sick leave, paid health insurance and private pensions.

The first two decades of the post-war era were dominated by institutional economists and scholars from sociology, political science, law, labour history, and psychology who united around a common desire to better understand and regulate labour management relations. In 1946 strike levels in the USA reached their historic peak. Concern over the escalating labour-management conflict led a number of state legislatures to create new multidisciplinary schools or centres of

industrial relations in leading universities such as Cornell, Wisconsin, Illinois, Michigan State, Rutgers and the University of California at both Berkeley and UCLA. In 1947 a new scholarly professional association, the Industrial Relations Research Association, was created. This association continues today under the name of the Labor and Employment Relations Association.

Two sets of questions featured prominently in industrial relations research in decades following the Second World War: (*a*) how does collective bargaining work and (*b*) what are the effects of unions and collective bargaining on management, the workforce and the economy? A debate arose over whether political (that is, pressures from union members and the need for union leaders to match settlements achieved in closely aligned industries or occupations) (Ross 1948) or economic forces (Dunlop 1944) were the primary drivers of wage determination. While never fully resolved, the evidence suggested that both play roles – political forces are influential within a range but are limited by market conditions. A reformulation of the debate by one institutional economist suggested that bargaining power includes a mixture of political, economic and 'pure power' forces and that these should be incorporated into a more complete theory of wage determination under collective bargaining (Levinson 1968).

The growing presence and pressure of unions and collective bargaining from the 1930s through the 1950s exerted what one set of researchers called a shock effect on management. Personnel practices had become more professionalized and applied in more uniform fashion and management had to search for ways to improve productivity to recoup the higher wage costs resulting from collective bargaining (Slichter et al. 1960). Much of industrial relations research over this time period examined the dynamics of labour management relations and the causes of strikes and/or industrial peace (Golden and Parker 1955). Most of this work was carried out using qualitative case studies or historical studies of specific unions or of industrial relations in particular industries.

Dunlop (1958) criticized post-war industrial relations research for being characterized by too many facts chasing too little theory. He sought to correct this problem by proposing a general systems theory of industrial relations. He argued that the central task for industrial relations theory was to explain variations in the rules governing employment relations. These rules were set in interactions among three key actors – labour, management and government – and conditioned by external market, technological and societal forces. The system was bound together by what Dunlop argued was a shared ideology valuing democracy, respect for market forces and worker rights. Although Dunlop's framework never reached the level of being accepted as a general theory of industrial relations, it became the starting point for much of what constituted industrial relations research in the decades following publication of this important work.

Public Sector Unions and Collective Bargaining

Government employees were not covered under the National Labor Relations Act (NRLA) of 1935 and, with a few exceptions such as postal employees, remained largely non-union until the 1960s. In 1958 Wisconsin enacted the first of what would grow to be a surge of state legislation protecting state and local employees' right to unionize and engage in bargaining. By 1976 38 states had enacted similar statutes, employing various forms of mediation, fact-finding, and arbitration to resolve contract disputes in lieu of the right to strike. Only a handful of states provided public employees the right to strike and even in these cases police and firefighters were not given the right to strike. Federal employees were granted similar rights to negotiate over non-wage and benefit issues, first through an Executive Order enacted in 1962 and then through legislation enacted in 1978.

As a result, unionism among public employees grew from its minimal level prior to 1960 to reach its present level (in 2007) of approximately 37 per cent of all government employees. These developments produced a significant body of new research on public sector collective bargaining

throughout the 1960s and 1970s. Most of this work focused on the performance of mediation, fact-finding and arbitration as deterrents to strikes. The consensus findings of these studies is that arbitration has been successful in deterring strikes of public employees (Olson 1988). Other studies have focused on the effects of public sector unions and collective bargaining on wages and government budgets. The general findings of these studies are that unions can increase wages. Prior to 1980, estimates suggested the union effect was around five per cent. After 1980 it rose to 20 per cent for local government employees and ten per cent for federal employees (Gunderson 2007).

Internal Labour Markets

The study of labour market behaviour represents another longstanding strand of research in industrial relations, dating back to Commons's (1909) classic historical study of changes in labour and product markets of shoemakers. Throughout the 1940s and 1950s studies of the dynamics of external labour markets followed the institutional tradition by examining the development of industry and regional wage structures (Lester 1952; Rees and Shultz 1970).

Interest turned to the study of internal labour markets in the 1970s and thereafter by both economists (Doeringer and Piore 1972; Osterman 1984) and sociologists (Baron and Bielby 1980; Pfeiffer and Baron 1988). Internal labour markets refer to firm-level rules governing hiring and termination, arrangement of jobs into job ladders, compensation structures that link jobs, and access to and mobility of personnel within and across job ladders. The primary questions of interest in these studies is what substantive rules govern the organization of jobs and job ladders and what factors give rise to the development, continuity and decline of internal labour market rules and practices. There is more consensus over the factors giving rise to internal labour markets than to the degree to which are the causes of their decline. Internal labour markets arise as a function of pressure from unions, governments and tight labour markets (Jacoby 1985). Over time these rules gain sufficient acceptance to become norms that sustain them even in the face of changing conditions in the external labour market (Osterman 1984). A major topic of debate in emerged in the 1990s over whether, and if so why, internal labour markets are declining in importance as firms appear to be more willing to lay off workers, adjust compensation to external market signals and hire more workers from outside the firm rather than train and promote current employees (Cappelli 1999; Jacoby 1999). There is no conclusive outcome to this debate. Micro firm-level studies tend to find more significant changes in firm-level rules and employment practices and outcomes while macro labour market studies tend to observe modest reductions in employee tenure (for men but not for women). The relative consensus is that norms governing layoff decisions and internal wage structures have led leading firms to be less reluctant to lay off hourly and managerial employees and more willing to allow their internal wage structures to become more disparate or unequal.

Resurgence of the Basic Disciplines

In the 1960s and 1970s the disciplines and methodologies from which industrial relations researchers drew became more quantitative as econometric and psychometric tools advanced, micro data-sets on labour market behaviour became more readily available, and computer power became more readily accessible. The vast majority of newly trained labour economists moved away from institutional analysis in favour of drawing propositions from neoclassical economics that could be tested with econometric methods. Studies of individual labour market behaviour grew and studies of collective behaviour, where data were less available, declined. Research on discrimination, mobility, labour supply, returns to education, and human capital flourished while the study of unions and collective bargaining declined.

The exception to the shift away from unions and collective bargaining was the use of econometrics to estimate the impact of unions on

relative wages of individuals (Lewis 1963). The consensus estimates of these studies were that private sector unions raised wages of their members relative to comparable non-members between 10 and 15 per cent. These estimates rose to 15 to 22 per cent in the 1970s (Kochan and Helfman 1981). Unions also were shown to have positive effects on other outcomes such as health and pension coverage, wage inequality, productivity, worker retention and satisfaction with wages (Bennett and Kaufman 2007). Unions have negative effects on firm profits and satisfaction with non-wage outcomes (such as satisfaction with job content) (Kochan and Helfman 1981; Freeman and Medoff 1984).

The development of human capital theory (Becker 1975) further encouraged the movement of labour economics back into the mainstream of the economics discipline and away from its institutional orientation. Becker's work stimulated others (Lazear 1998) to apply economic analysis to personnel decisions and practice. The study of alternative forms of incentive compensation and their effects on motivation and performance lies at the heart of personnel economics. A paradox appears to exist: the empirical evidence documents the economic value of incentive compensation to the firm, while use of individual incentives has not grown and in some countries appears to be in decline. Explaining this paradox requires consideration of the social context and other institutional forces that seek to reduce competition among workers and enhance social cohesion at work. Personnel economics' models of incentive compensation, therefore, need to be supplemented with sociological theories of group norms and other institutional factors that shape wage determination in contemporary organizations.

The same movement back to their mother disciplines could be observed by the 1970s in the work of psychologists and sociologists studying work and employment issues. Models of motivation, job satisfaction, work performance, turnover, and other aspects of individual attitudes and behaviours, based most often on survey, laboratory, or other data-sets assembled by these researchers, became the dominant topics and methodologies. This has given rise to the more applied field of human resource management research. Human resource management combines analysis of firm-level personnel functions (selection, compensation, performance appraisal, and so forth) with analysis of the links between human resource strategies and individual or organizational performance (Dyer 1984; Schuler and Jackson 1987). Most of this work adopted the normative premises of the human relations and scientific management schools rather than those of industrial relations. Thus they focused on how to manage employees through the use of modern personnel and human resource practices and strategies to overcome any sources of conflict in the employment relationship and to foster firm performance.

1980s: A Time of Transformation

The 1980s proved to be a watershed decade for both the study and practice of industrial relations. A central debate arose over whether reductions in real and nominal wage and other changes observed in collective bargaining were simply temporary adjustments to the deep recession of 1981–3 or signalled a more permanent structural shift in the wage determination process and in industrial relations more generally. Few today doubt that the wage determination and industrial relations practices shifted in fundamental ways in the 1980s by reducing the power of the strike threat, and weakening unions in general. Strike rates (measured in percentage of contract negotiations that involve a strike or percentage of annual work hours lost to strikes) have declined precipitously to the point they are no longer reported by government agencies.

The confluence of the deep recession, increased international competition and a shift to a conservative government in the United States under President Ronald Reagan unleashed a set of changes that created a set of anomalies for much of postwar industrial relations theory and empirical research. Management became more openly hostile and aggressive in avoiding new union organizing, moving operations from union to non-union workplaces. Management replaced

unions as the driving force in shaping the process and outcomes of collective bargaining. Nominal wage reductions were negotiated in many employment contracts. New approaches to work organization and employee participation challenged traditional job structures and labour management relations. These developments led to an expanded model of industrial relations that emphasized how the choices made by management in particular (but labour and government as well) in structuring relations at the workplace, in collective bargaining or personnel policies and in high-level business/ competitive strategies shape employment relationships and outcomes (Kochan et al. 1986).

Analysis of how these choices played out and affected outcomes featured significantly in industrial relations research throughout the 1980s and 1990s. Researchers began to assess the effects of different combinations of employment practices on firm performance, reflecting the systems' perspective of industrial relations and the emerging emphasis on complementary practices in personnel economics (Milgrom and Roberts 1992). By the end of the 20th century the evidence suggested that flexible work systems and employee involvement in production and workplace decisions served as positive complements to investments in technology and training, produced significant improvements in productivity and service quality (Ichniowski et al. 1996; Appelbaum et al. 2000). The theory and evidence suggested a high-wage, high-productivity equilibrium was possible in sectors as diverse as manufacturing, airlines, health care and financial services. Yet these 'high performance' work systems did not diffuse naturally across the economy, in part because of the costs of transitioning from more traditional practices, and in part because they competed with a low-wage, low-cost equilibrium. These two competing models of human resource practice and industrial relations compete with each other across most industries and occupations in the USA and other countries. A central theoretical and policy question in the field today focuses on whether a highwage, high-productivity equilibrium can be sustained in the face of low-wage, low- cost competition in domestic *and* international labour and product markets, and, if so, how to best encourage adoption of these strategies.

Policy Debates

Concerns over public policy rose in parallel to these theoretical and empirical developments. The central proposition driving policy debates was that the changes in the workforce, nature of work, and the economy had outpaced adaptations in public policies, institutions, and practices in employment relations and that this gap was imposing costs on workers and the economy (Osterman et al. 2001). Efforts to build consensus on changes needed in labour and employment policies consistently failed from the late 1970s to the 1990s (Kochan 1995). The result is that the field of industrial relations has come full circle to where it began in the early years of the 20th century when Commons and his students documented the mismatch between policies and institutions and workplace relations as the economy transitioned from its agrarian base to a manufacturing base. Today the mismatch is playing out on a global rather than a domestic scale, and therefore the theoretical, institution building, and public policy challenges are broader and perhaps more complex than ever before. As yet, however, there is little public or political support for comprehensive reforms of labour and employment policies. Consistent with the history of policy changes in the United States, it will likely take a significant crisis, combined with a major shift in political power, to achieve a change in policy.

Rebirth of Sociological Studies of Work and Labour Markets

While sociologists have studied various aspects of work, employment and careers throughout the 20th century (Hughes 1958; Barley and Kunda 2001), since the 1980s there has been a significant growth in interest in these topics among sociologists, who now label their work as the study of economic sociology.

Economic sociologists implicitly (and sometimes explicitly) seek to counter purely economic models of labour-market behaviour by demonstrating that individual and organizational decisions reflect the social and institutional structures in which they are embedded. Much of this work examines how networks of workers and/or organizations influence labour market behaviour. An early study in this tradition (Granovetter 1974) documented how networks affect access to job opportunities. Later studies have shown networks to be important in influencing migration (Portes and Sensenbrenner 1993; Saxenian et al. 2002), promotions and upward mobility (Burt 1992), and cooperative relationships among firms in industrial regions (Putnam 1973; Piore and Sabel 1984; Locke 1995). These studies extend the institutional tradition of industrial relations by drawing more heavily on classical sociological theories of Weber (1962), Durkheim (1893) and Selznick (1984).

From Industrial Relations to Work and Employment Relations

By the beginning of the 21st century many scholars began to recognize that the term 'industrial relations' had become increasingly problematic as a label of the study of people at work. The majority of the workforce is employed in services, not manufacturing. Thus many researchers and university programmes have gradually changed the labels used to describe their field of enquiry and/or teaching from industrial relations to work and employment relations, human resource management, work and organizational relations, and a variety of other terms. At the same time, more scholars from traditional disciplines of sociology, political science, economics and social psychology have taken up the study of work and employment issues, which has led to an expansion of the field and to a new round of competition among these different disciplines for influence in shaping the future study and practice of employment relations.

The research questions that are most central to this field today reflect two interrelated realities: (*a*) globalization of economic activity, and (*b*) the importance of knowledge and innovation in structuring work and shaping economic outcomes. Globalization and changes in technology have increased the mobility of capital, work, and workers thereby weakening the influence of national laws, institutions, and norms in shaping employment relationships and outcomes. Once again, today as in the Commons era of the early 20th century, wages and labour costs are under intense competition, only this time more labour markets are international in scope.

The increased ease of locating work and expansion of trade across national borders affects a wide range of work and employment issues and outcomes. Globalization has been associated with, among other things, changes in the distribution of wages and profits, growth in income inequality, and greater and more widely distributed job insecurity. Within firms, globalization of production and supply chains diffuses responsibility for employment decisions and policies, blurring the traditional distinction between employers and employees. All these effects are being subjected to intense analysis, debate, measurement of the direction and magnitude of their effects, and debate over how to adapt policies and institutions to cope with them. These international and organizational developments also make it more difficult to regulate employment relations with national laws and firm-centred rules and policies.

These developments have also generated a debate over the appropriate goals of the modern corporation and its role in society and as an employer. Since the early 1980s the view that firms exist primarily or even solely to maximize shareholder value has dominated academic and public discourse. This view is now being challenged. Blair and Stout (1999) offer a critique of the view that firms exist solely or primarily to maximize shareholder wealth and instead propose a team production theory of the firm. In their view the appropriate underlying view is that the firm should maximize the total value of wealth produced for all the constituents that supply resources and add value to the organization. Human capital plays a central role in this theory since workers

contribute and put at risk their human capital by joining and staying with a given firm. The longer workers stay with a given firm, the higher the costs of losing their job. Thus, like those who invest and put at risk their financial capital, workers are residual risk bearers should the firm fail.

The outcome of this debate could have important consequences for the design of institutions of worker voice in employment relationships. Since 1935 American labour law has taken as a guiding premise that employees should be allowed to bargain over wages, hours, and working conditions and that management should remain free to make strategic business decisions on its own. If by investing their human capital employees become a residual risk bearer similar to financial investors, then there is no logical basis for excluding workers from a voice in strategic decisions and corporate governance. Thus the study of industrial relations has expanded to engage issues of corporate strategy and governance and theories of the firm.

The field has also expanded in response to changes in the relationships between work and family–personal life. Work and family life were tightly linked in the preindustrial agrarian economy because they were co-located (families lived and worked on the farm) and men, women, and children all contributed to the production process. With the growth of the industrial economy came a clearer division of labour and physical separation in work and family life. The male breadwinner emerged as the prototypical worker, with the assumption that he had a wife at home attending to family responsibilities. With the growth in the labour-force participation of women from the 1960s onward and the slowdown in the growth of real wages, working hours have both been spread more evenly between men and women, and particularly between mothers and fathers. This once again increases the interdependence of work and family life and calls for changes in workplace and human resource practices to provide flexibility in hours and career options for women and men. Thus work and family issues have become an important topic of research and policy analysis within the field of work and employment relations (Bailyn 2006; Kossak 2006; Drago 2007).

International Studies

The study of work and employment relations across the world parallels most of the trends observed in the USA. Throughout much of the 20th century, studies of labour movements and labour conflict dominated both country-specific research and international comparisons of industrial relations systems. In the 1960s a debate arose over whether technological changes and increasing economic interdependencies would lead to a convergence in employment systems and practices or whether differences observed across countries would endure because of the influence of national culture and other institutional forces (Kerr et al. 1960). This debate continues today, although researchers have shifted to more micro level (industry, occupational and regional) comparisons to sort out forces leading to convergence and divergence in employment relationships (Katz and Darbashire 2000; Bamber et al. 2004). Moreover, researchers active in the field of international industrial relations (Kaufman 2004) are actively analysing and debating most of the issues and developments discussed in this article in countries across the globe.

Historical Parallel

In the USA and Britain, the field of work and employment issues of industrial relations have come full circle to their origins. As in the first two decades of the 20th century, contemporary researchers are driven by a broad proposition that the nature of the economy, workforce, the nature of work and its relationship to other institutions such as family life have all changed dramatically while public policies and institutions remain tailored to a fading industrial-based economy. The gap between policies and institutions and the contemporary realities of work and family life lie at the heart of the tensions and pressures building up in workplaces in America and,

increasingly, across the world. The central task of work and employment researchers today, as for their industrial relations forefathers, is to conduct research and policy analysis that prepares for the day that the political forces align to make it possible to begin the updating and modernization process.

See Also

- ▶ Collective Bargaining
- ▶ Commons, John Rogers (1862–1945)
- ▶ Human Capital
- ▶ Institutional Economics
- ▶ Slichter, Sumner Huber (1892–1959)
- ▶ Social Contract
- ▶ Strikes
- ▶ Taylorism
- ▶ Unemployment

Bibliography

Appelbaum, E., T. Bailey, P. Berg, and A. Kalleberg. 2000. *Manufacturing advantage*. Ithaca: Cornell University/ILR Press.

Bailyn, L. 2006. *Breaking the mold*. 2nd ed. Ithaca: Cornell/ILR Press.

Bamber, G., R. Lansbury, and N. Wailes. 2004. *International and comparative employment relations*. London: Sage.

Barbash, J. 1984. *The elements of industrial relations*. Madison: University of Wisconsin Press.

Barley, S., and G. Kunda. 2001. Bringing work back in. *Organization Science* 12: 76–95.

Baron, J., and W. Bielby. 1980. Bringing the firms back in: Stratification, segmentation, and the organization of work. *American Sociological Review* 45: 737–765.

Becker, G. 1975. *Human capital: A theoretical and empirical analysis*. New York: Columbia University Press.

Bennett, J., and B. Kaufman. 2007. *What do unions do: A twenty-year perspective*. New Brunswick: Transaction Press.

Blair, M., and L. Stout. 1999. A team production theory of corporate law. *Virginia Law Review* 85: 247–328.

Budd, J. 2004. *Employment relations with a human face*. Ithaca: Cornell University/ILR Press.

Burt, R. 1992. *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Cappelli, P. 1999. Career jobs are dead. *California Management Review* 42: 146–167.

Clegg, H. 1970. *The system of industrial relations in Great Britain*. London: Blackwell.

Commons, J. 1909. American shoemakers, 1648–1895. *Quarterly Journal of Economics* 24: 39–98.

Commons, J. 1934. *Institutional economics*. New York: Macmillan.

Doeringer, P., and M. Piore. 1972. *Internal labor markets and manpower analysis*. Lexington: D.C. Heath.

Drago, R. 2007. *Striking a balance*. Boston: Dollars & Sense.

Dunlop, J. 1944. *Wage determination under trade unions*. New York: Macmillan.

Dunlop, J.T. 1958. *Industrial relations systems*. New York: McGraw Hill.

Durkheim, E. 1893. *The division of labor in society*, 1984. New York: Free Press.

Dyer, L. 1984. Studying human resource strategy: An approach and agenda. *Industrial Relations* 23: 155–169.

Fox, A. 1971. *A sociology of work and industry*. London: Collier MacMillan.

Freeman, R., and J. Medoff. 1984. *What do unions do?* New York: Basic Books.

Golden, C., and V. Parker. 1955. *Causes of industrial peace*. New York: Harper and Row.

Granovetter, M. 1974. *Getting a job: A study of contacts and careers*. Cambridge, MA: Harvard University Press.

Gunderson, M. 2007. Two faces of union voice in the public sector. In *What do unions do: A twenty-year perspective*, ed. J. Bennet and B. Kaufman. New Brunswick: Transaction Press.

Hughes, E. 1958. *Men and their work*. Glencoe: Free Press.

Hyman, R. 1975. *Industrial relations: A Marxist introduction*. London: Macmillan.

Ichniowski, C., T. Kochan, D. Levine, C. Olson, and G. Strauss. 1996. What works at work: Overview and assessment? *Industrial Relations* 35: 299–333.

Jacoby, S. 1985. *Employing bureaucracy*. New York: Columbia University Press.

Jacoby, S. 1991. *Masters to managers*. New York: Columbia University Press.

Jacoby, S. 1999. Are career jobs headed for extinction? *California Management Review* 42: 123–145.

Katz, H., and O. Darbashire. 2000. *Converging divergences*. Ithaca: Cornell/ILR Press.

Kaufman, B. 1993. *The origins and evolution of the field of industrial relations*. Ithaca: Cornell University/ILR Press.

Kaufman, B. 2004. *The global evolution of industrial relations*. Geneva: International Labour Organization.

Kerr, C., J. Dunlop, F. Harbison, and C. Myers. 1960. *Industrialism and industrial man*. Cambridge, MA: Harvard University Press.

Kochan, T. 1980. *Industrial relations: From theory to policy and practice*. Homewood: Irwin.

Kochan, T. 1995. Using the Dunlop report for mutual gain. *Industrial Relations* 34: 350–366.

Kochan, T., and D. Helfman. 1981. The effects of collective bargaining on economic and behavioral job outcomes. In *Research in labor economics*, ed. R. Ehrenberg. Greenwich: JAI Press.

I

Kochan, T., H. Katz, and R. McKersie. 1986. *The transformation of american industrial relations*. New York: McGraw Hill.
Kossak, E. 2006. Work and family in America: Growing tensions between employment policy and a transformed workforce. In *America at work*, ed. E. Lawler and J. O'Toole. New York: Palgrave Macmillan.
Kuhn, T. 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Lazear, E. 1998. *Personnel economics for managers*. New York: Wiley.
Lester, R. 1952. A range theory of wage differentials. *Industrial and Labor Relations Review* 5: 433–450.
Levinson, H. 1968. Wage determination under collective bargaining. In *Collective bargaining*, ed. A. Flanders. London: Palgrave Books.
Lewis, H. 1963. *Unions and relative wages in the United States*. Chicago: University of Chicago Press.
Locke, R. 1995. *Remaking the Italian economy.* Ithaca: Cornell University Press.
Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.
Marx, K. 1849. Wages, labor capital. Reprint. In *The Marx-Engels reader*, ed. R. Tucker. New York: Norton, 1972.
Meltz, N. 1989. Industrial relations: Balancing efficiency and equity. In *Theories and concepts of comparative industrial relations*, ed. J. Barbash and K. Barbash. Columbia: University of South Carolina Press.
Milgrom, P., and J. Roberts. 1992. *Economics, organizations, and management*. Upper Saddle River: Prentice Hall.
Olson, C. 1988. Dispute resolution in the public sector. In *Public sector bargaining*, ed. B. Aaron, J. Najita, and J. Stern, 2nd ed. Washington, DC: Bureau of National Affairs.
Osterman, P. 1984. *Internal labor markets*. Cambridge, MA: MIT Press.
Osterman, P., T. Kochan, R. Locke, and M. Piore. 2001. *Working in America: A blueprint for the new labor market*. Cambridge, MA: MIT Press.
Pfeiffer, J., and J. Baron. 1988. Taking workers back out: Recent trends in the structuring of employment. In *Research on organizational behavior*, ed. B. Staw and L. Cummings, vol. 10. Greenwich: JAI Press.
Piore, M., and C. Sabel. 1984. *The second industrial divide*. New York: Basic Books.
Portes, A., and J. Sensenbrenner. 1993. Embeddedness and immigration: Notes on the social determinants of economic action. *American Journal of Sociology* 98: 1320–1350.
Putnam, R. 1973. *The beliefs of politicians*. New Haven: Yale University Press.
Rees, A., and G. Shultz. 1970. *Workers and wages in an urban labor market*. Chicago: University of Chicago Press.
Roethlisberger, F., and W. Dickson. 1939. *Management and the worker*. Cambridge, MA: Harvard University Press.
Ross, A. 1948. *Trade union wage policy.* Berkeley: University of California Press.
Saxenian, A., Y. Motoyama, X. Quan, and D. Wittenborn. 2002. *Local and global networks of immigrant professionals in Silicon Valley.* San Francisco: Public Policy Institute of California.
Schuler, R., and S. Jackson. 1987. Linking competitive strategies with human resource management practices. *Academy of Management Executive* 1: 207–219.
Selznick, P. 1984. *Leadership in administration*. Berkeley: University of California Press.
Slichter, S., R. Livernash, and J. Healy. 1960. *The impact of collective bargaining on management*. Washington, DC: Brookings Institution.
Taylor, F. 1895. A piece rate system, being a step toward partial solution of the labor problem. *Transactions* 16: 856–883.
Walton, R., and R. McKersie. 1965. *A behavioral theory of labor negotiations*. New York: McGraw Hill.
Webb, S., and B. Webb. 1894. *The history of trade unions*. London: Longmans, Green & Co.
Webb, S., and B. Webb. 1897. *Industrial democracy.* London: Longmans, Green & Co.
Weber, M. 1962. *Basic concepts in sociology.* New York: Philosophical Library.
Wright, P.M., and G.C. McMahan. 1992. Theoretical perspectives for human resource management. *Journal of Management* 18: 295–320.

Industrial Revolution

Gregory Clark

Abstract

The term 'Industrial Revolution' has come to mean two very different things: first, the transformation the British economy experienced between 1760 and 1850, to become the first modern industrialized, fast-growing economy; second, the general switch between the pre-industrial world of slow technological advance, high fertility and little human capital to the modern world of rapid efficiency gains, low fertility and large investments in human capital. Modern economists' theories of this second worldwide transition have proved difficult to reconcile with the details of Britain's transition.

Keywords

Child mortality; Demographic transition; Education; Enlightenment; Family planning; Fertility; Fertility–income relationship; Human capital; Industrial Revolution; Innovation; Malthusian economy; Population growth; Real wage rates; Skill premium; Technical change; Total factor productivity

JEL Classifications

N1

The Industrial Revolution is an ambiguous term, freighted with multiple meanings, interpreted differently by different writers. First, it describes the extraordinary transformation the British economy experienced between 1760 and 1850. In these years Britain moved from being a largely self-sufficient, self-sustaining, and still principally agrarian society, to being an economy where a substantial fraction of food, raw materials and energy was imported, or mined from the earth as coal, and where the great majority of the population was engaged in industry and commerce. But second, and more importantly, it has come to mean the general move in the world economy in about 1800 from the pre-industrial economy, which experienced extremely low rates of efficiency growth, to the modern economy, where efficiency growth is rapid and persistent. That shift from low rates of efficiency advance to rapid rates had nothing inherently to do with industry or industrialization. Efficiency advance in agriculture has been as rapid as in the rest of the economy since 1800. So for the more general use of the term 'Industrial Revolution' the 'industrial' component is a misnomer, but a misnomer that we have to live with.

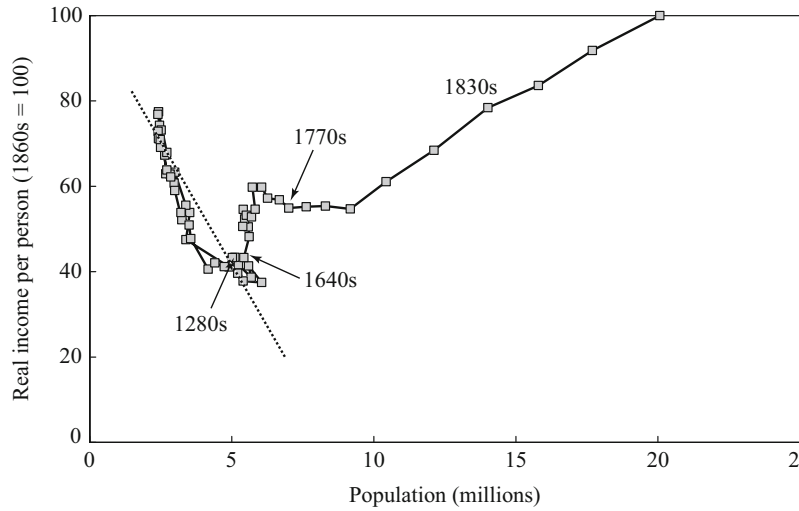
The Industrial Revolution of the Historians

The 'Industrial Revolution' more traditionally describes a specific period in British history, most commonly taken as 1760 to 1850. In 1760 Britain was a prosperous but still heavily agrarian economy, with half the labour force employed in agriculture. Foreign trade was insubstantial. Britain was largely self-sufficient in staple foods. The main imports were Mediterranean or tropical products such as sugar and spices, wines, raisins, coffee and tea. The main export was woollen cloth produced by domestic weavers or handloom workshops. London was already a huge city with over 750,000 inhabitants, but the other towns in England circa 1760 were mostly small. The next biggest city was Bristol with only 50,000 people. Travel and communication were slow and costly. The road system was poorly maintained, and there were few canals.

By 1850 the share of the population employed in agriculture in Britain had dropped to less than a quarter. Staple foods and raw materials such as timber had become major imports. Exports were dominated by factory-produced textiles, but included a whole range of manufactured goods and even substantial amounts of coal. The urban population had grown enormously. Manchester, for example, had grown from about 20,000 in 1770 to over 300,000 by 1851. London had nearly 2.4 million by 1851, more than 13 per cent of English people, and was the largest city in the world. The road system had greatly improved, and alongside the roads there were now about 2000 miles of canals and improved river navigations, as well as more than 5000 miles of the new railways.

Rapid population growth accompanied the change in occupational structure, location and trade patterns. The English population grew from seven million in the 1770s to 19 million by the 1850s. Periods of population growth earlier in English history, as in the 13th and the 16th centuries, were associated with declining living standards. The Industrial Revolution represented a sharp break with this past. For the first time living standards improved even as the population swelled. Figure 1 shows the real wage of building workers vis-à-vis the English population from 1250 to 1850. The unusual character of experience in the Industrial Revolution era is clear.

Between 1760 and 1850 England experienced what was cumulatively profound economic change, though the actual rate of change for



Industrial Revolution, Fig. 1 Real building workers' day wages vis-à-vis population by decade, 1280–1849. *Note*: The line summarizing the trade-off between population and real wages for the pre-industrial era is fitted using the data from 1280–9 to 1590–9 (*Source*: Clark 2005b, Fig. 5)

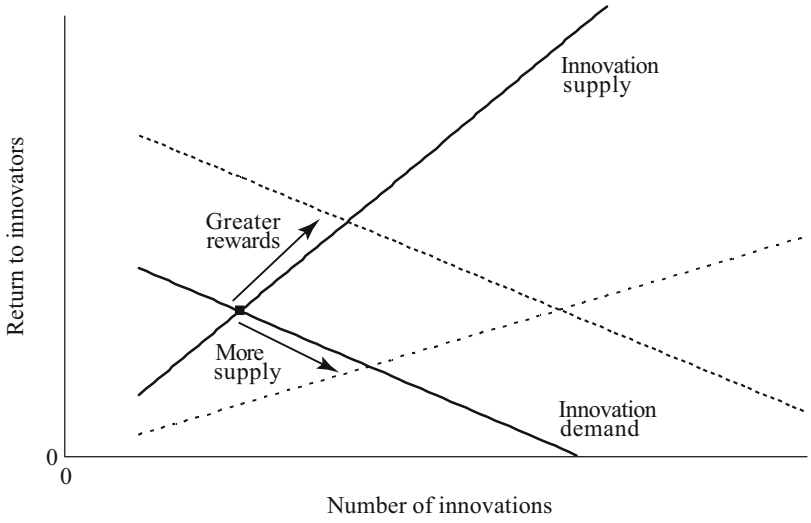
most measures of the economy such as gross output per person or the fraction of the population employed in agriculture was by modern standards very slow. Indeed, the changes were so slow that many economists writing in this period – such as Adam Smith, Thomas Malthus and David Ricardo – had little comprehension of the fundamental break from the past that was occurring.

The recent consensus has been that the immediate cause of the Industrial Revolution was the dramatic increase in efficiency in a minority of the economy: yarn and cloth production, iron and steel making, and rail transport. Most of the economy, including surprisingly the coal industry, saw little technological advance (Clark and Jacks 2007). Textiles alone explain perhaps 60 per cent of all measured technological advance from 1760 to 1850. The concentration of technological advance in textiles, aligned with the move of production there into factories, explains why the general move around 1800 towards economies with faster technological advance came to be labelled the 'Industrial Revolution'.

In textiles we see a whole series of innovations, especially from the 1760s onwards, which transformed the industry. These innovations had no direct connection with the scientific advances of the previous 150 years and were indeed mainly made by artisans and craftsmen with no formal scientific training. Nor were the new production processes in these industries particularly capital-using. Water and steam powered textile mills were modest in their capital requirements compared with later innovations like the railways, but also compared with existing industries like agriculture. The demands of these mills were mainly for unskilled labour. Tending the new spinning and weaving machines did not require literacy, and involved skills fully mastered within a year of employment. Thus the Industrial Revolution in the first instance did not involve great investments in either physical or human capital.

The question of why England first experienced the Industrial Revolution, and why only in the 1760s, has occupied the energies of an enormous number of historians and economists. There has been an intense debate on the features of the British economy in 1760 that precipitated the break from the past. Generations of economic historians have thrown themselves at the problem, like waves of infantry in the First World War going over the top of the trenches. Their explanations, however, have generally fared no better than the average First World War soldier when tested against the history of England in these years.

Putative explanations of the Industrial Revolution can be separated into those based on the supply of or the demand for innovations, as portrayed in Fig. 2. Some emphasize greater returns to innovation as inducing more innovation, others a greater supply of innovators.



Industrial Revolution, Fig. 2 Demand and supply interpretations of the Industrial Revolution

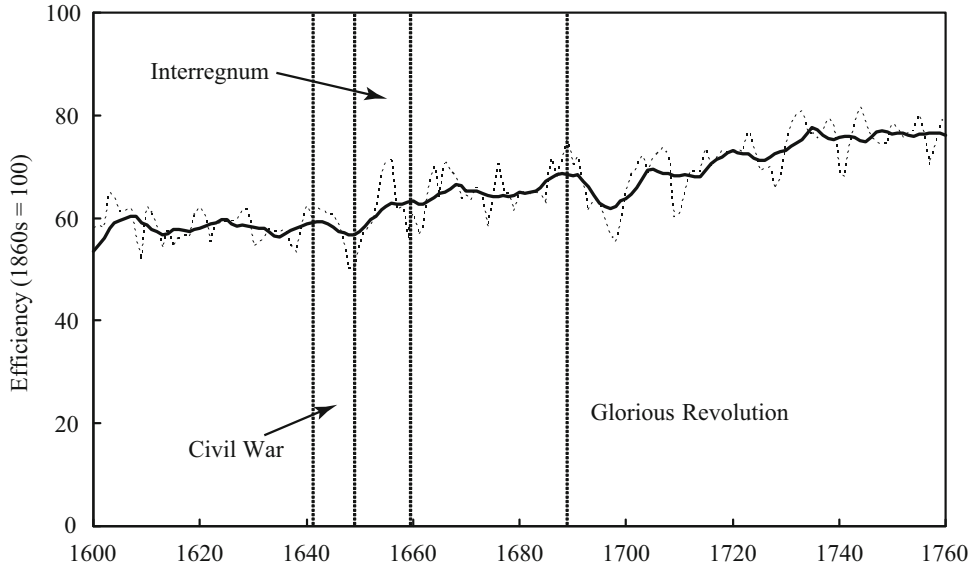
Much attention has been given, for example, to the institutional changes that preceded the English Industrial Revolution, and raised the benefits to innovation. Douglass North and Barry Weingast proclaimed the Glorious Revolution of 1688–9, which established the institutional framework of the modern British state with a figurehead monarch and control by an elected parliament, as the key precondition for economic growth (North and Weingast 1989). The development of a government restrained from seizing the profits of investors increased the expected returns to investment in general in the economy.

There are numerous problems with this identification. The gap between the institutional changes and the onset of the Industrial Revolution is a generous 80 years or so. In those 80 years there were was no speed up in the rate of efficiency advance in the economy, as Fig. 3 shows. The efficiency of the economy, known also as the total factor productivity (TFP), is the amount of output delivered per unit of input of capital, labour and land. From 1689 to 1760 the English economy had efficiency growth rates no faster than those of the 'bad' days of the old regime in 1600–89, when England experienced considerable political turmoil.

Also, contemporary economic actors seem to have attached no importance to the political changes of 1688–9. Gross rates of return on capital in the private economy, for example, did not decline, as would be expected if the new regime had ushered in more secure property rights (Clark 1996). Finally, societies such as that of England had most of the institutional prerequisites of modern growth – stable politics, free markets, factor mobility, and low taxation – hundreds of years before any growth appeared (Clark 2007, ch. 8).

Kenneth Pomeranz has argued that the Industrial Revolution was triggered in England in the 1760s, and not in other sophisticated societies such as China, because of the accidents of coal and colonies (Pomeranz 2000). The chance location of coal fields in England, and the ability of North America to supply massive imports of raw materials liberated England from the energy and raw material constraints that had limited growth before in the self-sustaining organic pre-industrial economy. But the concentration of growth in cotton textiles, an industry that was present also in Japan and China by 1800, where water power could supply all the energy required, suggests that the elements Pomeranz concentrates on were actually peripheral to the Industrial Revolution (Clark and Jacks 2007).

Other economists, such as Joel Mokyr, have argued alternately that the root cause of the Industrial Revolution was an increased supply of innovation, promoted by the Enlightenment, the intellectual movement which swept Europe in



Industrial Revolution, Fig. 3 Efficiency level of the English economy, 1600–1860. *Notes*: The figure shows the estimated efficiency of the English economy by year (*dotted line*) and as an 11-year moving average (*solid line*) (*Source*: Clark 2007, Fig. 12.6)

the 18th century (Mokyr 2005). Mokyr shows that while the Enlightenment was an important intellectual movement in many European countries such as France it was a particularly prominent part of intellectual life in England. And if we look at many other measures – literacy, numeracy, publications – England was becoming a more intellectually sophisticated society in the years leading up to the Industrial Revolution at all levels of the society. But Mokyr offers no account of why this intellectual movement should have taken hold in England in particular, and only in the 18th century.

The Industrial Revolution of the Economists

From a broader perspective, the Industrial Revolution that brought us from the static pre-industrial economy to the modern dynamic economy is characterized by a three key features.

Most important is the appearance of persistent total factor productivity growth. Such growth occurs when output rises faster than the measured inputs. Thus if y is output per worker hour, k capital services per worker hour, and z land services per worker hour, and A the level of efficiency (TFP) of the economy, A grows at the rate

$$g_A = g_y - a \cdot g_k - c \cdot g_z$$

where g denotes a growth rate, and a and c are the shares of capital and land in total factor costs. Since 1850 in the most successful economies TFP has grown at one per cent or more per year. Before 1800, over extended periods, even for successful economies TFP grew at rates of 0.01–0.1 per cent per year.

We can estimate TFP growth before 1800 using population. On average before 1800 output per worker-hour, y, did not rise (see the ▶ Malthusian Economy). In this case we can simplify the equation above. In such a static economy, labour hours L will be proportionate to population N. Since the land area is fixed

$$g_z = -gL = -gN.$$

Similarly income per capita was constant over the long run. On the assumption that the rate of return on capital did not change, capital per person would have been constant, so that,

$$g_k = 0.$$

Substituting both these relations into the basic equation above implies that for the pre-industrial world the growth rate of efficiency over the long run was just

$$g_A = c \cdot g_N.$$

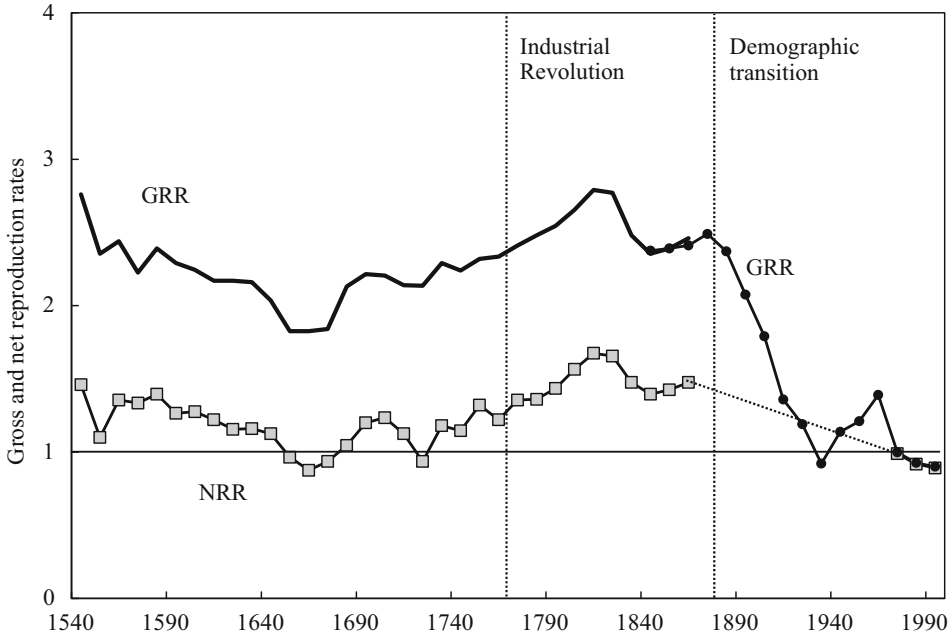
Thus long-run technological advance at a world scale before 1800 is proportionate to long-run population growth, as Kremer (1993) pointed out. Since plagues or disorder can result in wages departing from the long-run equilibrium, this calculation serves only for the long run. Table 1 shows the details. For the world as a whole there is no long period before 1700 when the rate of technological advance even exceeds 0.1 per cent per year.

The second general feature of the broader Industrial Revolution has been declining fertility, measured as births per woman. English women, for example, average five births each all the way from the 1540s to the 1890s. Figure 4 shows the

Industrial Revolution, Table 1 Growth rate of world population and TFP before 1800

Year	Population (millions)	Population growth rate (%)	Technology growth rate(%)
130,000 BC	0.1	–	–
10,000 BC	7	0.004	0.001
1 AD	300	0.038	0.009
1000 AD	310	0.003	0.001
1250 AD	400	0.102	0.025
1500 AD	490	0.081	0.020
1750 AD	770	0.181	0.045

Source: Clark (2007, Table 7.1)



Industrial Revolution, Fig. 4 English fertility history, 1540–2000. *Notes*: *GRR* gross reproduction rate. *NRR* net reproduction rate. The data for the years after 1837 is for the whole population. Before 1837 it is from a sample of parishes (*Sources*: Clark 2005a, Fig. 2)

gross reproduction rate (GRR), the number of daughters born per woman living to the age of 50, by decade in England from the 1540s to 1990s. The 'demographic transition' to modern fertility rates in Europe and North America, except for France, began only in the 1880s. By 2000 English women gave birth on average to fewer than two children.

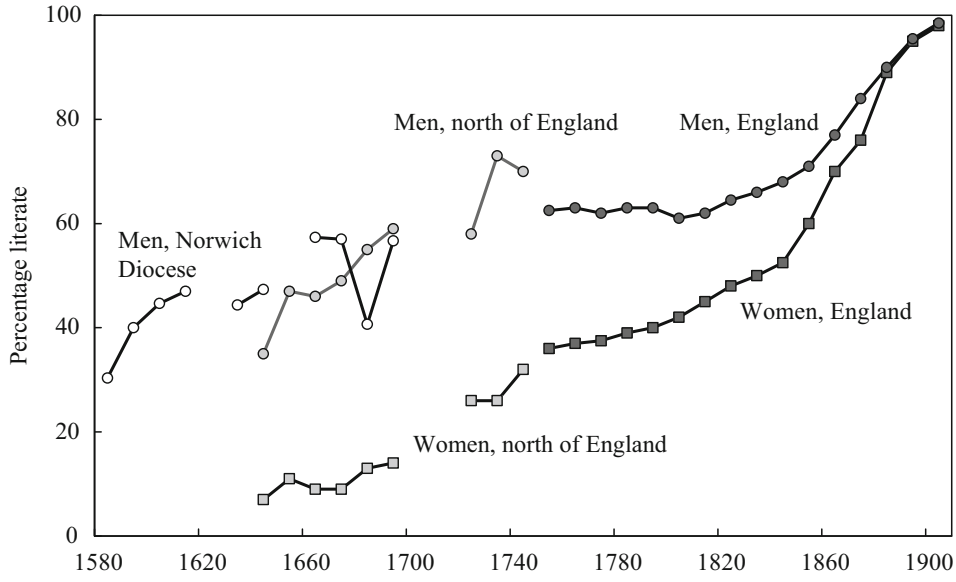
Since pre-industrial child mortality rates were high, however, the net reproduction rate (NRR), the number of daughters the average woman gave birth to over her lifetime, fell much less in the modern world than in the pre-industrial era. Figure 4 shows also the NRR for England. England in 1540–1800 had an unusually high NRR for pre-industrial society, where this number would normally be just slightly above 1. Note that the GRR and NRR both rose in England in the course of the classic Industrial Revolution.

The decline in gross fertility after the 1880s was crucial in allowing enhanced efficiency in the economy to translate into higher incomes. Had this not happened, so that population growth would have been much more rapid, then the share of payments to land as a factor, c, would not have declined so rapidly and might even have increased. Then in the first equation above the increase of population per acre would have been faster, and its weight greater, leading to a greater drag on income growth.

The third key feature of the transition to the modern world has been an increase in human capital per person, investments in education and training. In most pre- industrial societies the mass of the population was illiterate and innumerate. Along with the Industrial Revolution came a transition to a society where the implied value of human capital is nearly as great as for physical capital.

English education levels increased over the Industrial Revolution years. Figure 5 shows a measure of basic literacy, the fraction of men and women signing their names on witness statements or marriage registers. However, if one compares Fig. 5 with Fig. 4 there appears to be no connection between changes in literacy rates and changes in fertility: the fertility transition in England occurred after the attainment of mass literacy.

The coincidence of these three great changes in societies – technological advance, declining birth rates and increased education – has led economists in recent years to attempt theories of the broader Industrial Revolution that unify these



Industrial Revolution, Fig. 5 Literacy in England, 1580–1920 (*Source*: Clark 2005a, Fig. 3)

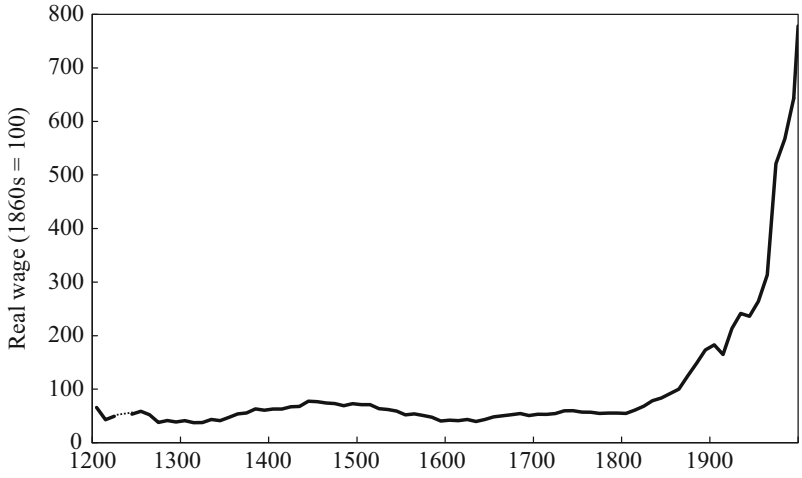
elements (Becker et al. 1990; Galor 2005; Galor and Moav 2002; Galor and Weil 2000; Lucas 2002). These theories, however, face formidable obstacles in reconciling themselves to the facts of the Industrial Revolution in England.

One method of unification would posit the technological advances as primary, and have the income gains from these spur both lower fertility and more investment in human capital. In the years of the demographic transition in both the USA and in Europe between 1880 and 1920, higher-income families were the first to reduce fertility (Clark 2007, Table 14.5; Jones and Tertilt 2006, pp. 23–7). Indeed, Larry Jones and Michele Tertilt conclude that, for female birth cohorts in the USA between 1828 and 1958, income explains most of the decline in gross fertility. Figure 6, for example, shows the hourly real wage of building workers in England from 1200 to 2000. After the 1860s real wages begin to rise rapidly, and after the 1860s fertility declined substantially. In the modern world there is a strong negative fertility–income relationship across countries.

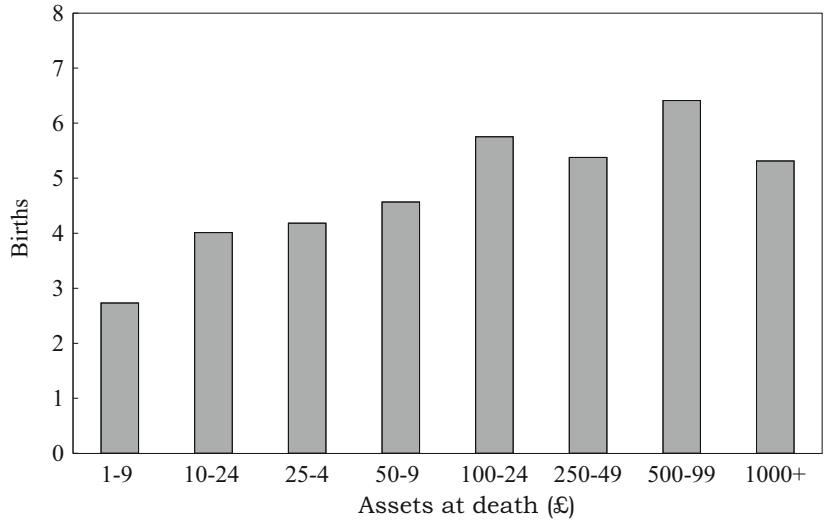
The problem with explaining the fertility transition through income is that all plausible models of population regulation before 1800 depend on a *positive* association between fertility and income. Empirical information on pre-industrial fertility and income is rare. But in pre-industrial England we get an insight into the connection through evidence from the wills of male testators (Clark and Hamilton 2006). Connecting information on assets at death to parish records reveals the average numbers of births per testator for each bequest class. Figure 7 shows that a man leaving less than £25 at death would typically father fewer than four children, while one with assets of more than £1000, six children. Thus in pre-industrial England there was a positive association between income and both gross and net fertility over a wide range of incomes. This stands in sharp contrast to the association in the modern world.

This positive association between fertility and income became negative in the period of demographic transition. But in current high-income, low-fertility societies there seems to be only the most modest negative association between income and fertility. A recent study of female fertility found on average little association between household income and fertility, measured as the numbers of children present in the households of married women aged 30–42, for 1980 and 2000, for the six Organisation for Economic Cooperation and Development (OECD) countries (Dickmann 2003, Table 2). The income–fertility relationship within societies has changed dramatically over time.

All this makes constructing a link between fertility and income challenging. Why does fertility increase with income in the pre-industrial world? Authors who have addressed this have concentrated on explaining the association for incomes close to subsistence level. Galor and



Industrial Revolution, Fig. 6 Real day wages of English building workers, 1200–2000 (*Source*: Clark 2005b, Fig. 1)



Industrial Revolution, Fig. 7 Births by assets of testator, 1585–1636 (*Source*: Clark 2007, Fig. 4.3)

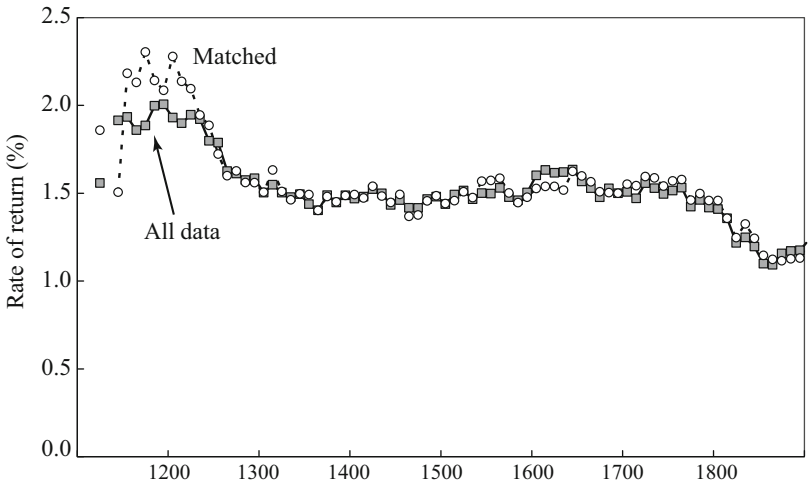
Weil (2000) and Galor and Moav (2002) assume a minimum consumption level that parents must achieve before producing children. Lucas (2002) assumes children require a minimum consumption transfer. We see in Fig. 7, however, that the richest families in pre-industrial England, people who would have high incomes even by the standards of 1900 showed high gross fertility rates.

The third problem with using income to explain declining family size is that, as Fig. 6 shows, we cannot explain rising human capital in the years prior to Industrial Revolution through income gains. Human capital gains preceded the income gains of the Industrial Revolution. Finally, as noted above, we still lack any institutional or other explanation for the transition towards higher rates of efficiency advance after 1800.

Another mechanism that might explain both the rise in human capital and the decline in fertility and the Industrial Revolution would be an increase in the premium paid for human capital in the Industrial Revolution era. In most settled pre-industrial economies the bulk of labour demand was for agricultural work, where levels of human capital were low. In such an economy, it is argued, parents would favour quantity over 'quality' in children.

However, for this explanation is to be compatible with individual incentives, the return from investments in human capital before the Industrial Revolution has to be low. In England, and in a variety of other pre-industrial economies, rewards to human capital were higher than in the modern economy. We have, for example, the skill premium in the building industry: the ratio of the wages of craftsmen to building labourers. Figure 8 shows the wages of craftsmen relative to labourers in England by decade from 1200. The period 1600–1900, when literacy rates increased markedly, featured a near constant skill premium. When fertility rates fell after 1800 it was in a labour market where the premium for skills was also declining markedly. Thus gross fertility is highest where the premium for skills in the labour market is greatest. A demand interpretation of fertility decline, on its own, will not work either in England or as a general explanation of the fertility transition.

Since the expansion of human capital first occurred when the return to human capital was constant, the gains of human capital in the Industrial Revolution era had to involve significant supply shifts. Galor and Moav (2002) posit that the supply shift was created by Darwinian competition in the pre-industrial economy between families with different tastes for child 'quality'. In the Malthusian world each family can have a NRR only slightly above 1. But 'high-quality'



Industrial Revolution, Fig. 8 The skill premium for building workers, 1200–2000 (*Source*: Clark 2005b, Fig. 2)

families do better. High- quality types produce offspring who, because of their greater human capital and hence higher incomes, have more children. Thus, when incomes are close to subsistence, but only then, they out-produce the 'low-quality' types. There should be an inverse U-shape of fertility with income. Figure 7, however, is inconsistent with this proposed mechanism. Even the richest in pre-industrial England show the highest gross and net fertility rates.

Clark (2007), however, argues that more general Darwinian selection mechanisms in the pre-industrial era could explain the move to more human capital, and the greater supply of innovations in the Industrial Revolution. Just as people were shaping economies, the economy of the pre-industrial era was shaping people, at the least culturally, perhaps also genetically. The Neolithic revolution created agrarian societies that were just as capital intensive as the modern world. At least in England, the emergence of such an institutionally stable, capital-intensive economic system created a society that rewarded middle-class values with reproductive success, generation after generation. This selection process was accompanied by changes in characteristics of the pre-industrial economy that owe much to the population displaying more middle-class preferences. Interest rates fell, murder rates declined, work hours increased, the taste for violence declined, and numeracy and literacy spread even to the lower reaches of society. These selection mechanisms thus provide an economic underpinning to the intellectual developments such as the Enlightenment of the 18th century that Mokyr identifies as a key background to the Industrial Revolution in England.

But such an explanation for the onset of the Industrial Revolution, which emphasizes the greater fertility of the rich in the pre-industrial era, leaves declining fertility after 1880 as a conundrum. If the economic system prior to the Industrial Revolution selected those with a tendency to use higher incomes to achieve greater net fertility, why did all this change in the 1880s? There are several possible explanations.

One is that the desired number of children per married couple is actually independent of income, and was always for just two or three surviving children. But to ensure a completed family size of even two children in the high-mortality environment of the Malthusian era required six or more births. For example, in pre- industrial England where 60 per cent of children died before adulthood, to ensure a 90 per cent chance of getting a surviving son would require giving birth to seven children. Nearly 40 per cent of the poorest married men leaving wills in 17th century England had no surviving son. Even among the richest married men nearly one-fifth left no son. The average rich man left four children because some families had large numbers of surviving children. Hence the absence of any sign of fertility control by richer families in pre-industrial England may

stem largely from the uncertainties of child survival in the Malthusian era. This may have led to an unwillingness on the part of all families to limit births. As the fraction of children surviving increased in the late 19th century, even risk-averse families could afford to begin limiting births.

In the late 19th century child mortality in England had fallen substantially from the levels of the 18th century, and the rate of that decline was strongly correlated with income. For families living in homes with ten or more rooms only 13 per cent of children failed to reach the age of 15, while for those in one room still 47 per cent of children failed to reach that age (Clark 2007, p. 00). Thus the lower gross fertility of high-income groups at the end of the 19th century translates into a more muted decline in net fertility. And these groups faced a substantially reduced variance in family size outcomes compared with low-income groups.

Another possible element in the decline of fertility since the Industrial Revolution is the increased social status of women. Men may well have had greater desire for children in pre-industrial society than women. Women, not men, bore the very real health risks of pregnancy, and did most of the work involved in bringing up the children. But typically men had a much more powerful position within the family. Thus women may always have desired smaller numbers of surviving children than men, but have been able to effect those desires only in the late 19th century.

Women's relative status and voice was clearly increasing in the late 19th century in England, when literacy rates for women had advanced to near equality with those of men. Women had gained access to universities by 1869, enhanced property rights within marriage by 1882, votes in local elections in 1894, and finally a vote in national elections in 1918. The gain in the relative status and voice of women proceeded most rapidly among higher-income groups.

These assumptions could explain why net fertility falls after the late 19th century – even though in cross section in the 16th century – and in 2000 there is either a positive connection between income and net fertility or no connection. They could also explain why the demographic transition appeared first in the higher socio- economic status groups, so that net fertility is negatively related to income in the transition period.

See Also

- ▶ Historical Demography
- ▶ Malthusian Economy

Bibliography

Becker, G., K. Murphy, and R. Tamura. 1990. Human capital, fertility and economic growth. *Journal of Political Economy* 98: S12–S37.

Clark, G. 1996. The political foundations of modern economic growth: England, 1540–1800. *Journal of Interdisciplinary History* 26: 563–588.

Clark, G. 2005a. Human capital, fertility, and the industrial revolution. *Journal of the European Economic Association* 3: 505–515.

Clark, G. 2005b. The condition of the working-class in England, 1209–2004. *Journal of Political Economy* 113: 1307–1340.

Clark, G. 2007. *A farewell to alms: A brief economic history of the world*. Princeton: Princeton University Press.

Clark, G., and G. Hamilton. 2006. Survival of the richest: The Malthusian mechanism in pre-industrial England. *Journal of Economic History* 66: 707–736.

Clark, G., and D. Jacks. 2007. Coal and the Industrial Revolution, 1700–1869. *European Review of Economic History* 11: 39–72.

Crafts, N.F.R. 1985. *British economic growth during the Industrial Revolution*. New York: Oxford University Press.

Crafts, N.F.R., and C.K. Harley. 1992. Output growth and the Industrial Revolution: A restatement of the Crafts–Harley view. *Economic History Review* 45: 703–730.

Dickmann, N. 2003. *Fertility and family income on the move: An international comparison over 20 years*, Working paper no. 360. Syracuse: Maxwell School of Citizenship and Public Affairs, Syracuse University..

Galor, O. 2005. From stagnation to growth: Unified growth theory. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Galor, O., and O. Moav. 2002. Natural selection and the origin of economic growth. *Quarterly Journal of Economics* 117: 1133–1191.

Galor, O., and D.N. Weil. 2000. Population, technology and growth: From Malthusian stagnation to the demographic transition and beyond. *American Economic Review* 90: 806–828.

Jones, L.E., and Tertilt, M. 2006. *An economic history of fertility in the U.S. 1826–1960*, Working paper 12796. Cambridge, MA: NBER.

Kremer, M. 1993. Population growth and technological change: One million BC to 1990. *Quarterly Journal of Economics* 108: 681–716.
Lucas, R.E. 2002. The industrial revolution: Past and future. In *Lectures on economic growth*, ed. R.E. Lucas. Cambridge, MA: Harvard University Press.
McCloskey, D.N. 1981. The industrial revolution 1780–1860: a survey. In *The economic history of Britain since 1700*, ed. R. Floud and D. McCloskey, vol. 1. Cambridge: Cambridge University Press.
Mokyr, J. 1999. Editor's introduction: The new economic history and the Industrial Revolution. In *The British Industrial Revolution: An economic perspective*, ed. J. Mokyr, 2nd ed. Boulder: Westview Press.
Mokyr, J. 2005. The intellectual origins of modern economic growth. *Journal of Economic History* 65: 285–351.
North, D.C., and B. Weingast. 1989. Constitutions and commitment: Evolution of institutions governing public choice in seventeenth century England. *Journal of Economic History* 49: 803–832.
Pomeranz, K. 2000. *The great divergence: China, Europe, and the making of the modern world economy.* Princeton: Princeton University Press.

Industrialization

Amiya Kumar Bagchi

Industrialization is a process. The following are essential characteristics of an unambiguous industrialization process. First, the proportion of the national (or territorial) income derived from manufacturing activities and from secondary industry in general goes up, except perhaps for cyclical interruptions. Secondly, the proportion of the working population engaged in manufacturing and secondary industry in general also shows a rising trend. While these two ratios are increasing, the income per head of the population also goes up except again for temporary interruptions (Datta 1952; Kuznets 1966, 1971; Sutcliffe 1971). There are cases in which the per capita income goes up, income derived from secondary industry per head of the population also goes up, but there may be little growth either in the proportion of income derived from the secondary sector or in the ratio of the working force engaged in that sector. Such cases, except when they are observed for a highly developed country, not only make the unambiguous labelling of the process of development as industrialization difficult; they also pose questions regarding the sustainability of the process that has been observed.

Other characteristics are also often associated with industrialization or a more general process of what Kuznets has called 'modern economic growth' (Kuznets 1966, ch. 1). These include a narrowing and ultimate closing of the gap between productivity per head in the secondary sector and in the primary sector (that is, agriculture, forestry and fishing), continual changes in the methods of production, the fashioning of new products, rise in the proportion of population living in towns, changes in the relative ratios of expenditures on capital formation and consumption and so on.

Most of these associated characteristics were derived from the experience of Great Britain, or more narrowly, England and Wales, which was the first country to industrialize. That experience has remained unique in many ways. But since England was the original centre for diffusion of the economic and technical changes associated with the industrialization process, it is important to understand what happened in that country.

At least since the days of Karl Marx, England has been known as the first country in which feudalism broke down, and capitalism brought the economy under its sway (Dobb 1946). This meant that all means of production came to be owned by a small group of property-owners called capitalists, and the rest of the working people became free wage-workers who earned their livelihood by selling their labour power to the capitalists (Dobb 1946). It has been claimed that while serfdom broke down all over western Europe, England was the only country where a group of landlords managed to concentrate most of the land in their hands and prevent the consolidation of a free peasantry which could be used by an absolutist state to defeat the rise of capitalist agriculture (Moore 1967, ch. 1; Brenner 1976). It has been further claimed that economic individualism which has been taken as the hallmark of the motivation of an entrepreneur in capitalist society, goes back in England to the 12th–13th centuries,

so that capitalism went through a long process of birth in the first industrializing nation (Macfarlane 1978, chs 5–7). By the time of the first industrial revolution, England was a society in which the abiding interest of the rulers was to make money from agriculture, trade and industry, and in which the rulers were prepared rationally to order the affairs of the state so as to enable the entrepreneurs to conquer foreign countries and markets, by the force of arms if need be and had the financial and military might to carry out those plans. England had also become the leader in trade and finance among the countries of western Europe, after the decline of Amsterdam (Braudel 1984).

The English industrial revolution is traditionally associated with the rise of machine-based industry powered by steam (Marx 1887, chs XIV and XV; Mantoux 1928). Certainly the classic age of British dominance of world industry, which is roughly the period from the end of the Napoleonic wars up to 1870, was characterized by the conquest of production methods by machines with moving parts of iron and steel, powered by steam, and operated by scores or even hundreds of operatives concentrated in single factories. However, what is becoming apparent is that for practically the whole of the eighteenth century, traditional techniques and materials (such as wood), and traditional sources of power such as muscles of men, women and children, animals, and water and wind, were responsible for the growth and spread of factory industry (Musson 1972; Von Tunzelmann 1978; Crafts 1985).

The experience of England lends credence to the postulation of a stage of 'industrialization before industrialization' or 'proto-industrialization' (Mendels 1972). This has been defined as 'the development of rural regions in which a large part of the population lived entirely or to a considerable extent from industrial mass production for inter-regional and international markets' (Kriedte et al. 1981, p. 6). The growth of industry in England was spearheaded by an explosion in the development of cotton spinning; and the cotton mills which utilized the new spinning machines sought out suitable sources of water power and labour – mostly in the rural areas or small towns. Steam engines were an element in the industrial revolution, but they did not come into their own as the major prime movers in manufacturing industry until perhaps the second quarter of the nineteenth century.

In England, cotton textiles were a relatively new industry; and they grew at first by redressing the balance of labour power needed in traditional spinning methods, so that no major displacement of labour took place within the system of proto-industrialization, in the 18th century. But machine spinning stimulated handloom production, and handloom weavers were pauperized even in England when powerlooms displaced handlooms (Bythell 1969). In other countries, where traditional handicrafts were displaced by the new machine-made fabrics, and because of political or internal social factors they were not replaced, or not replaced quickly enough, by any considerable growth of machine industry, pauperization and de-industrialization were widespread and in some cases they became endemic phenomena (Bagchi 1976; Kriedte 1981). So Ricardo's worries about the possible employment-displacing effects of machinery were justified after all (Ricardo 1821, ch. 31; see also Hicks 1969). But in Britain, continued growth of external trade and the coming of the railway age helped in overcompensating the labour-displacing effects. Not all countries had the same advantages.

The proto-industrialized order was succeeded in England by the system of machine manufacture perhaps because the former faced its severest crisis there: social relations there had already been transformed in a fully capitalist mould by the time smallscale manufacture reached its fullest development. Developments in science, technology and statecraft almost certainly helped resolve the crisis in favour of a higher stage of industrial development.

The fact that England had a decisive lead in the use of machine manufacture and steam power, and had formal or informal colonies where she could ignore barriers erected by the USA or Continental European countries made her the supreme industrial nation of the world for almost three quarters of a century (cf. Robinson 1954).

Once the revolution in textiles and steam power had been pioneered in England it could,

however, be diffused to other countries, provided the latter possessed suitable political and social conditions. It is on the basis of the timing, speed and social mechanism of diffusion of the industrial revolution that we can distinguish three clusters of countries which have gone through an unambiguous process of industrialization. The first is the cluster of countries on both sides of the North Atlantic seaboard and overseas colonies with populations of predominantly European origin; the second consists of Japan and the four islands of industrialization in the Far East, viz., South Korea, Taiwan, Hong Kong and Singapore, and the third is the cluster of socialist countries led by the Soviet Union. The rest of the world are still struggling, with only varying degrees of success, to get a sustained process of industrialization going (Bagchi 1982).

The English industrial revolution was, to start with, very much a matter of textiles; it was only in the 19th century that it affected other industries, especially iron and steel and mechanical engineering in general, on a large scale. The uniqueness of England, with all the advantages of a first start (Robinson 1954) allowed her to expand her markets overseas in an almost unrestrained manner until the USA and other western European countries expanded their home production, not only of textiles, but also of other manufactures, often behind walls of protection against the English manufactures. The west European industrialization was helped very much by the nearness of England: from Britain flowed information about the new inventions, machines, men and capital, although there was for a time an attempt to restrict the exports of new machinery from England (Landes 1965). Capital flows from England and to a lesser extent from France, were particularly important in supporting the movement of European populations to the USA, Canada, Australia, South Africa, New Zealand and Argentina (Kuznets 1971; Bagchi 1972; Edelstein 1982).

Yet despite more active support by politically independent governments, the spread of industrialization to western Europe took a surprisingly long time to get going (Lewis 1978, chs 7 and 8; Crafts 1985, ch. 3). One set of reasons had to do with political, social and structural factors. The French needed a major revolution before the bourgeoisie could take possession of the state apparatus. Even then, the entrenchment of peasant agriculture in the countryside probably delayed the full conversion of the primary sector to capitalist relations. In other countries, even the 1848 revolution did not complete the process of capitalist take-over. Associated with these lags went the fact that by English standards, too high a proportion of population continued to depend on agriculture and a large gap between agricultural and industrial productivity continued to persist down to the eve of World War I. Such political and social lags, of course, even more impeded the process of industrialization in the countries of central, southern and eastern Europe down to the period between the two world wars. We will have to pay separate attention to the case where the logjam in the process of industrialization was only broken with the Bolshevik Revolution, with most other countries of eastern Europe following after World War II (Berend and Ranki 1982).

As the process of industrialization spread, the supply of importable technologies and the financial requirements for implementing such technologies both increased. According to one estimate, gross domestic investment as a proportion of GDP in Great Britain increased from around 4 per cent in 1700 to 5.7 per cent in 1760, 7.9 per cent in 1801 and 11.7 per cent in 1831 (Crafts 1983), and remained between 10 and 12 per cent between 1831 and 1860 (Feinstein 1978, p. 91). By contrast in countries such as Germany, Sweden or Denmark the rate of investment in their phase of industrialization (after 1860) often reached 15 per cent and more of GDP. In the USA the social preconditions for industrialization were much more favourable than in most European countries, and the export of capital from Europe considerably aided her industrialization process until she in turn became a creditor nation around the turn of the 19th century.

The latecomers among the western European countries, and Japan on the other side of the world, used state intervention on a much wider scale and much more purposively than Britain did. This intervention did not take the same form in all countries: in a country such as Germany,

financing of industry was far more widely supported by the state and by new instruments of finance created for the purpose than, say, in Italy. It is doubtful whether a general pattern of successful state intervention to overcome economic backwardness can be discerned from the historical experience as has sometimes been claimed (Gerschenkron 1962). What can be asserted is that state intervention in industry was much more likely to succeed in countries where capitalist relations had advanced far than where intervention from the top was used as a substitute for social change which might upset the balance of class forces among the rulers (cf. Berend and Ranki 1982).

The example of Russia is especially instructive in showing the limits of state action in a society where capitalist relations had taken root only to an imperfect degree. In Russia serfdom had been consciously introduced in the 17th century, and the system bore particularly heavily on regions producing grain which was a major export of eastern European lands. The so-called village communes (*obschina*) produced both agricultural products and handicrafts. Beginning around the 1830s modern machinery was employed in the processing of beet sugar and in the spinning of cotton yarn. Even after the abolition of serfdom in 1861, handicrafts remained predominant (Crisp 1978), but later on, the system of domestic production and production by handicrafts became more and more integrated into the system of capitalist production (Lenin 1898). It was only in the 1880s that large-scale industry employing modern machines experienced an accelerated growth in Russia (Lyaschenko 1949; Crisp 1978).

The development of modern industry and capitalism in general in Russia gave rise to a vigorous debate which still has contemporary relevance in many countries of the third world. Some of the Russian Populists (*Narodniks*) contended that the development of capitalist industry was impossible in a backward country such as Russia. They argued that modern machine-based industry destroys handicrafts and small peasant agriculture, the incomes of people dependent on them consequently shrink, and thus modern industry faces a severe – indeed insurmountable – realization problem. Countering this argument, Lenin pointed out that capitalism created its own markets by converting goods produced within a household or barter economy into tradable commodities and by continually generating new methods of production. The latter in their turn create demands for new equipment and materials (Lenin 1897, 1899). Lenin did not deny that capitalism needed foreign markets. But at that stage he attributed the need not to the impossibility of realizing the surplus value but to intercapitalist competition and the continuous drive of capital towards expansion. In the process of discussing the analytical issues involved Lenin enunciated a law of development of capital, namely, that 'constant capital grows faster than variable capital, that is to say, an ever larger share of newly-formed capital is turned into that department of the social economy which produces means of production' (Lenin 1897, pp. 155–6).

While markets expanded in Russia with state support for development of railways and warrelated industries, the process of industrialization before the Revolution of 1917 remained ridden with numerous contradictions. Before the Stolypin reforms (which were initiated after the abortive revolution of 1905) the spread of individual ownership in agriculture was held up by numerous restrictions on peasant mobility and on the transferability of land. Even after the Stolypin reforms (or reaction) landlords' social and economic power continued to limit the development of capitalism in agriculture (see, e.g., Lenin 1912). A substantial proportion of growth in the industrial capital stock was financed by foreign banks and foreign entrepreneurs (McKay 1970). Large-scale industry was regionally and sectorally concentrated (Portal 1965) and the proportion of the working force engaged in industry (including construction) was only 9 per cent in 1913; it was only after the Bolshevik Revolution and the implementation of the two Five Year Plans that there was a decisive change in the occupational structure. The proportion of the working force engaged in industry and construction climbed to 23 per cent in 1940 and 39 per cent in 1979; correspondingly the proportion engaged

in the agriculture and forestry declined from 75 per cent in 1913 to 54 per cent in 1940 and 21 per cent in 1979 (Sarkisyants 1977, p. 180; see also Kuznets 1966, p. 107).

In Japan, the course and the pattern of industrialization differed considerably from the sequence witnessed in western Europe and the USA, and also from that followed in socialist countries. Under Tokugawa rule, Japan was characterized by what has been called 'centralized feudalism' (Ohkawa 1978, p. 140) with the *shogun* exercising supreme power through the *daimyos* and a rigid hierarchy going down to the village level. But the increasing use of money for the payment of taxes, countrywide transactions in money required to support the *daimyos'* and their retainers' expenditures in their travels to the capital and back, and the increasing indebtedness of many *daimyos* to merchants enhanced the power of the latter. The merchants' ambitions, the peasants' discontent and the frustrations of many of the feudal lords in the face of the increasing threat posed by the military and technological advance of the Western powers ultimately led to the end of the shogunate and the restoration of the Meiji emperor. The fierce nationalism bred among the nobility under the isolation enforced on the country earlier by the shogunate led them to define their objectives in the image of the activities of the Western imperialist powers (Beasley 1963; Norman 1943; Smith 1961).

While abolishing many of the privileges of the warrior class the new Japanese rulers held on to the rigid rules of hierarchal control descending from the emperor through the nobility and the higher ranks of merchants to the village headmen, and down to the peasants working in the fields. The rigid subjugation of family members, especially of women, to the patriarch and the use of communal ties to enforce authoritarian rule continued unabated, and was adapted to the requirements of modern industry (Morishima 1982). A high level of land taxes imposed on the peasantry financed much of the economic growth in Japan which accelerated from the 1880s. Young women, more or less bonded to the factories by their fathers or other family heads provided cheap labour. The first steps in the industrialization process were taken under the guidance of the state which built or financed shipyards, telegraph lines, railways and armament works (Lockwood 1968). The actual pace-setter in the industrialization process, in Japan as in Britain, was textiles, and for a long time, handicraft methods continued to be used alongside of machine methods in producing Japan's industrial goods. Silk, indemnities from foreign conquest, and exports of cotton yarn and cotton goods allowed Japan to do without much foreign investment in her drive towards industrialization. As in Britain, so in Japan, external markets and imperial conquest played an important role in the rise of modern industry (Lockwood 1968).

Japan's industrial growth was already impressive in its diversity and sophistication during the interwar years. But it is since World War II that her growth has surpassed earlier historical standards (Ohkawa 1978; Armstrong et al. 1984). The reserve army of labour in agriculture was finally exhausted there under the dual impact of land reforms imposed by the American occupation authorities and rates of industrial growth that often exceeded 15 per cent per year. Accompanying the Japanese growth was domination of trade and industry by a handful of giant conglomerates, the *zaibatsu*, giant firms and general trading corporations or *soga soshas*, acting in close collaboration with the Ministry of International Trade and Industry, and the subjugation of the labour movement to company objectives. It is these characteristics combined with a systematic exclusion of foreign capital from practically all fields that led observers to use the phrase 'Japan Inc.' to characterize the Japanese system of management. Japan eventually surpassed all capitalist countries except the USA in the value of her industrial production and in her technological advance.

While the countries on the two sides of the north Atlantic seaboard were industrializing and Japan was slowly emerging as a challenger to the industrial and political supremacy of the Western powers in the Far East, the majority of the people living in Asia, Africa and Latin America hardly experienced any positive process of industrialization. The movement of neither capital nor labour

favoured such a process in China, India, Egypt, Peru, Brazil or Mexico, even in the exceptional days of massive British investments overseas that enriched the USA, Australia or Canada with men or materials (Edelstein 1982; Davis and Huttenback 1985). Only a small fraction of foreign investment made by Britain and France went to the non-white, dependent colonies, or formally independent, but effectively dependent countries peopled by non-white populations. These investments went generally into plantations, mines, railways rather than manufacturing industries. While the British dominions such as Canada or Australia pursued their economic policies largely independently of metropolitan control and protected their nascent industries, India, Egypt or even China and Turkey were forced to pursue *laissez faire* policies under the pressure of the metropolitan powers. The small flows of foreign investment into the colonies were swamped in the case of India, West Indies or even Brazil by the outflow of capital to the metropolitan countries (and thence to their colonies of settlement) as political tribute, and profit on external trade, foreign exchange transactions or plantation and railway enterprises (Bagchi 1982, chs. 3 and 4).

Policies of free trade or state intervention in favour of metropolitan trade and industry generally led to a decline in handicrafts and domestic industry on a large scale in such countries as India, China and Turkey. This erosion of proto-industrial output and employment was only very inadequately compensated by the rise of modern industry. Colonial rule also led in many cases to the strengthening of ties of bondage of various kinds in the rural areas. When migration occurred on a large scale from these countries, it was often organized by the merchants from the metropolitan countries, and the migrants often entered into a semiservile condition in the plantations of Assam (India), Trinidad, Guiana, or mines of South Africa. The effective control of modern plantation and mining enterprises and many areas of trade, especially wholesale internal and external trade by merchants from metropolitan countries, policies of free trade and processes of de-industrialization retarded the development of an indigenous mercantile community in most of the dependent colonies and often delayed the onset of any process of industrialization until the 1950s.

In many Latin American countries industrialization was quickened in the 1930s as a result of import restriction policies forced on the governments by the deep depression, especially in primary commodity exports, and attendant balance of payments crises. Following on from this experience, many of them adopted industrialization as a strategy of development and the basic objective of planning. The Prebisch–Singer thesis of a secular decline in terms of trade of primary products provided the rationale for such a strategy in Latin America (Prebisch 1950; Singer 1950; Spraos 1982). Elsewhere, the success of the Soviet experiment provided an inspiration for planning.

However, after some initial successes, in most countries of the third world, the process of industrialization was caught up in multiple contradictions. In few countries were there land reforms conferring the right of ownership and control on the cultivating peasantry. This failure rendered the supply of food grains and other farm products inelastic and enabled the entrenched landlords and traders to speculate in these commodities. As a result, any stepping up of investment through governmental efforts soon met inflation barriers and balance of payments crises. The latter were aggravated by a tendency to import newer and newer consumer goods for the upper and the middle classes, by an inability to bargain from a position of strength with the suppliers of technology, and by the oversell practised by many of the aid-givers wanting to tie the loans or grants to purchases of goods from the donor country. In many Latin American countries, threats of social revolution were met by imposition of authoritarian regimes, generally with US connivance or assistance. The case of Chile where a popular government under the presidentship of Salvador Allende was replaced by a brutal military dictatorship is perhaps the most glaring example of this tendency, but Argentina, Brazil and Uruguay all fitted the same pattern. The primary commodities boom in the early 1970s, rise in oil prices in 1973–4 and 1978–9 along with the attraction provided to transnational corporations by explicit policies of wage repression and labour

regimentation boosted the rate of industrial growth in countries as widely dispersed as Brazil and Iran. However, in most of these countries, including some oilexporters such as Mexico and Nigeria, astronomically large external debts and debt servicing charges put a stop to most development efforts by the early 1980s.

A few economies in east Asia, more specifically the two enclaves of Hong Kong and Singapore, and the two medium-sized economies of South Korea and Taiwan went through a process of successful industrialization. In South Korea and Taiwan, radical land reforms, partly brought about through the defeat of the Japanese in 1945 who had been major land-holders in these two provinces, and partly imposed by the US authorities fearing a Communist revolution in emulation of the People's Republic of China, enormously speeded up the movement of trading capital into industry, increased the elasticity of supply of farm products and widened the market for basic consumer and producer goods. Chinese overseas capital had for a long time dominated trade and money-lending in many countries of east and south-east Asia. Communist take-over of mainland China drove out a sizeable section of big mercantile capital. The newly migrating and old Chinese overseas capital then turned to industrial investment in many of these countries. Increased US military activities in the region, attempted economic blockade of Communist China by the Western capitalist countries and the large expenditures attending US military aggression in Vietnam provided multiple opportunities to the traders and industrialists in the region for capital accumulation and expansion. Many Japanese, American and western European transnational corporations found Singapore, Hong Kong, Taiwan and South Korea useful as export platforms since these four economies provided the attraction of low wages, a disciplined (and regimented) labour force and privileged access to US, EEC and Japanese markets.

However, despite the fact that many Asian economies have continued to experience positive growth in a period of global recession, it cannot be said yet that the east Asian experience is catching or easily diffusible. Most of Africa is experiencing negative growth, and large clusters of population are caught there in the clutches of famine. Most Latin American economies are yet to get out of the debt trap. The only other countries which are still experiencing a positive process of industrialization to a greater or lesser extent are the socialist countries which have embraced some variant of Marxism as their guiding ideology. The share of industry in GNP rose steeply in most of these countries and often exceeded 40 per cent. The high rate of economic growth in these countries was financed by the confiscation of rent incomes from land, by the channelling of all surpluses into investment and by allowing only a moderate rise in real wages until an acceptable level of GNP was reached (cf. Ellman 1975; Lippit 1974). One feature that has distinguished socialist industrialization is that usually the share of services in GNP and employment has been lower than in most non-socialist economies. In a country such as China, the abnormally low share of services has been seen as a defect associated with the phase of extensive growth.

Most of the socialist countries are also now grappling with problems of lower productivity growth. Effective decentralization of planning processes, increased responsiveness to changes in relative scarcity and signalling of such changes through changes in relative prices and provision of adequate incentives to managers and workers have been generally seen as the answer to these problems. Increased imports of technology from the OECD countries and their effective absorption are also seen as part of the answer, but the successful pursuit of such strategies is intertwined with the issue of economic reforms on the one hand and geopolitical manoeuvring between the two blocks on the other.

One problem that will continue to bedevil industrialization strategies in most large third world countries for a long time is the very high ratio of the working population engaged in agriculture to the total working force. Even in a country such as China, which has experienced a trend rate of industrial growth of more than 10 per cent over the years since the Communist revolution in 1949, and where the share of industry in national income went up to 42.2 per cent in 1982,

agriculture and forestry continued to employ 71.6 per cent of the labour force in the same year (China 1983, pp. 24, 121). It is only some medium-sized economies with a high rate of industrial growth such as South Korea and Taiwan that have experienced any major shift in the population balance towards industrial employment.

The experience of the structural changes within the east Asian group of capitalist economies shows that under favourable circumstances, it is possible for the less industrialized economies to grow at high rates if there is a sustained shedding off of the lower-productivity sectors by the more advanced regions and the grafting of the shedded output on to the structures of the less developed economies (Yamazawa et al. 1983). The process is very similar to that observed in western Europe in the early part of the 19th century, except that the role of migration of population to countries outside the region (such as the USA in the case of western Europe) in easing population pressure has been minimal. But the roles of direct investment by Japanese and other OECD firms and of privileged access to extra-regional markets have been more significant than in the case of western Europe. (It could, of course, be argued that western European countries had a privileged access to markets in their dependent colonies.)

The general developments in the advanced capitalist bloc of countries (within which Japan occupies a unique position because of her maintenance of moderate to high rates of growth and near full employment and her large trade surpluses with most other countries), however, preclude the replication of the east Asian pattern in the rest of the third world. Most of them are afflicted by high rates of unemployment – exceeding levels witnessed since the end of the 1930s. Some countries such as the UK experienced an absolute decline in manufacturing (Singh 1977). These developments aggravated protectionism in these countries, thus creating barriers against the expansion of exports from the third world countries, while the OECD group of countries continued to constitute the biggest market for manufactured goods in the world. Developments in microelectronic technology posed major threats to the further expansion of labour-intensive textile products and clothing exports from the third world to the OECD countries (UNCTAD 1981). More generally, the spread of microelectronic technologies embracing whole branches of manufacture are threatening to remove many assembly operations which the OECD-based transnational corporations had earlier found it profitable to subcontract to the favoured export enclaves including the east Asian group of newly industrializing countries (Kaplinsky 1984).

Within the OECD group, the USA has become the biggest magnet for capital flows from all over the world. The high interest rate and large budget deficits maintained by the US government have forced most other OECD governments to pursue deflationary policies within their borders. There is little sign as yet that such trends will be reversed. The Japanese, who have run up large trade surpluses (exceeding US $40 billion) with the USA have proceeded to invest most of their export surplus in the US. Thus the diffusion that is borne on the backs of foreign investment within the order of capitalism has been severely hampered by these developments.

The only alternative that is left for most third world countries is to rely on building industries on the basis of domestic resources and domestic markets. But the guidelines laid down by the International Monetary Fund seeking to impose severely deflationary policies on most countries applying for its assistance in meeting their debt problems, the power exerted by OECD-based transnational corporations in effectively restricting the flow of technology, and the internal social structures in most of these countries blocking the spread of literacy and accrual of purchasing power to common people are likely to hamper the feeble efforts at industrialization on a self-reliant basis. The other path of industrialization, building on growing exports and expanding international investment flows, would appear also to be beset with dangerous pitfalls for most of the poor countries of the world. Thus the spread of industrialization in the near term to the poorer countries is likely to be very slow compared with the speed witnessed between, say, 1950 and 1978. At the other end of

the spectrum, in countries such as the USA and UK, services and finance have gained tremendously at the expense of manufacturing industry, and it is through the use of financial instruments as much as advanced technology in manufacturing (including armaments production) and services that the USA dominates the economies of most of the capitalist countries. But as Japan continues to forge ahead even in frontier technologies such as the mass production of semiconductor chips for use in the most advanced microelectronic processes (cf. Gregory 1985), a change in the balance within the capitalist order is very likely. In the meanwhile continued growth in the socialist world will also affect the global balance in manufacturing and economic power.

See Also

- ▶ Backwardness
- ▶ Dual Economies
- ▶ Gerschenkron, Alexander (1904–1978)
- ▶ Industrial Revolution
- ▶ Labour Surplus Economies
- ▶ Manufacturing and De-industrialization
- ▶ Mode of Production

Bibliography

Armstrong, P., A. Glyn, and J. Harrison. 1984. *Capitalism since World War II: The making and breaking of the great boom*. London: Fontana.

Bagchi, A.K. 1972. Some international foundations of capitalist growth and underdevelopment. *Economic and Political Weekly* 7(31–33): 1559–1570, Special Number.

Bagchi, A.K. 1976. De-industrialization in India in the nineteenth century: Some theoretical implications. *Journal of Development Studies* 12(2): 135–164.

Bagchi, A.K. 1982. *The political economy of underdevelopment*. Cambridge: Cambridge University Press.

Bairoch, P. 1975. *Economic development of the third world since 1900*. London: Methuen.

Beasley, W.G. 1963. *The modern history of Japan*. New York: Praeger.

Berend, I.T., and G. Ranki. 1982. *The European Periphery and Industrialization 1780–1914*. Cambridge: Cambridge University Press.

Blackaby, F. (ed.). 1979. *De-industrialization*. London: Heinemann.

Braudel, F. 1984. *The perspective of the world: Civilization & capitalism, 15th–18th century*. London: Collins.

Brenner, R. 1976. Agrarian class structure and economic development in pre-industrial Europe. *Past and Present* 70: 30–75.

Bythell, D. 1969. *The handloom weavers: A study in the English cotton industry during the industrial revolution*. Cambridge: Cambridge University Press.

China. 1983. *Statistical yearbook of China 1983*. Hong Kong: State Statistical Bureau PRC and Economic Information & Agency.

Crafts, N.F.R. 1983. British economic growth, 1700–1831: A review of the evidence. *Economic History Review* 36(2): 177–199.

Crafts, N.F.R. 1985. *British economic growth during the industrial revolution*. Oxford: Clarendon Press.

Crisp, O. 1978. Labour and industrialization in Russia. In Mathias and Postan (1978).

Datta, B. 1952. *The economics of industrialization*. Calcutta: World Press.

Davis, L., and R.A. Huttenback. 1985. The export of British finance, 1865–1914. *Journal of Imperial and Commonwealth History* 13(3): 28–76.

Dobb, M. 1946. *Studies in the development of capitalism*. London: Routledge & Kegan Paul.

Edelstein, M. 1982. *Overseas investment in the age of high imperialism: The United Kingdom 1850–1914*. London: Methuen.

Ellman, M. 1975. Did the agricultural surplus provide the resources for the increase in investment in the USSR during the first five year plan? *Economic Journal* 85: 844–863.

Feinstein, C.H. 1978. Capital formation in Great Britain. In Mathias and Postan (1978).

Gerschenkron, A. 1962. *Economic backwardness in historical perspective*. Cambridge, MA: Harvard University Press.

Gregory, G. 1985. Chip shop of the world. *New Scientist*, 15 August, 28–31.

Habakkuk, H.J., and M. Postan (eds.). 1965. *The Cambridge economic history of Europe*, The industrial revolution and after, Pts 1 and 2, vol. VI. Cambridge: Cambridge University Press.

Hicks, J. 1969. *A theory of economic history*. Oxford: Clarendon Press.

Hilton, R. (ed.). 1976. *The transition from feudalism to capitalism*. London: New Left Books.

Kaplinsky, R. 1984. The international context for industrialization in the coming decade. *Journal of Development Studies* 21(1): 75–96.

Kriedte, P. 1981. The origins, the agrarian context, and the conditions in the world market. In Kriedte, Medick, and Schlumbohm (1981).

Kriedte, P., H. Medick, and J. Schlumbohm. 1981. *Industrialization before industrialization: Rural industry before the genesis of capitalism*. Cambridge: Cambridge University Press.

Kuznets, S. 1966. *Modern economic growth: Rate, structure and spread*. New Haven: Yale University Press.

I

Kuznets, S. 1971. *Economic growth of nations: Total output and production structure*. Cambridge, MA: Harvard University Press.
Landes, D. 1965. Technological change and development in Western Europe 1750–1914. In Habakkuk and Postan (1965).
Lenin, V.I. 1897. A characterisation of economic romanticism (Sismondi, and our native Sismondists). Trans. from Russian in Lenin, *Collected works*, vol. 2. Moscow: Foreign Languages Publishing House, 1963.
Lenin, V.I. 1898. The handicraft census of 1894–95 in Perm Gubernia and general problems of handicraft industry. Trans. from Russian in Lenin, *Collected works*, vol. 2. Moscow: Foreign Languages Publishing House, 1963.
Lenin, V.I. 1899. *The development of capitalism in Russia*. Text of the 2nd edn of 1908, Trans. from Russian in Lenin, *Collected works*, vol. 3. Moscow: Progress Publishers, 1964.
Lenin, V.I. 1912. The last valve. Trans. from Russian in Lenin, *Collected works*, vol. 18. Moscow: Progress Publishers, 1968.
Lewis, W.A. 1978. *Growth and fluctuations 1870–1913*. London: Allen & Unwin.
Lippit, V.D. 1974. Land reform and economic development in China. *Chinese Economic Studies* 7(4): 3–181.
Lockwood, W.W. 1968. *The economic development of Japan*. Princeton: Princeton University Press.
Lyaschenko, P.T. 1949. *History of the national economy of Russia to the 1917 revolution*. London: Macmillan.
Macfarlane, A. 1978. *The origins of English individualism*. Oxford: Basil Blackwell.
Mantoux, P. 1928. *The industrial revolution in the eighteenth century*. London: Jonathan Cape.
Marx, K. 1867–94. *Das Kapital*. Trans. by S. Moore and E. Aveling as *Capital: A critical analysis of capitalist production*, vol. 1. Reprinted. Moscow: Foreign Languages Publishing House, n.d.
Mathias, P., and M.M. Postan (eds.). 1978. *The Cambridge economic history of Europe*. Vol. VII, *The industrial economies: Capital, labour and enterprise*, Pts 1 and 2. Cambridge: Cambridge University Press.
McKay, J.P. 1970. *Pioneers for profit: Foreign entrepreneurship and Russian industrialization*. Chicago: University of Chicago Press.
Mendels, F. 1972. Proto-industrialization: The first phase of the industrialization process. *Journal of Economic History* 32(1): 241–261.
Moore Jr., B. 1967. *Social origins of dictatorship and democracy: Land and peasant in the making of the modern world*. London: Allen Lane.
Morishima, M. 1982. *Why has Japan 'succeeded'?* Cambridge: Cambridge University Press.
Musson, A.E. (ed.). 1972. *Science, technology and economic growth in the eighteenth century*. London: Methuen.
Norman, E.H. 1943. *Soldier and peasant in Japan*. New York: Institute of Pacific Relations.
Ohkawa, K. 1978. Capital formation in Japan. In Mathias and Postan (1978).
Portal, R. 1965. The industrialization of Russia. In Habakkuk and Postan (1965), Pt 2.
Prebisch, R. 1950. *The economic development of Latin America and its principal problems*. New York: United Nations. Reprinted in *Economic Bulletin for Latin America* 7(1): 1–22 (1962).
Ricardo, D. 1821. *On the principles of political economy and taxation*. 3rd ed, reprinted in *The works and correspondence of David Ricardo*, Vol. I, ed. P. Sraffa with the collaboration of M.H. Dobb, Cambridge: Cambridge University Press, 1951.
Robinson, E.A.G. 1954. The changing structure of the British economy. *Economic Journal* 64: 443–461.
Sarkisyants, G.S. (ed.). 1977. *Soviet economy: Results and prospects*. Moscow: Progress Publishers.
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(2): 113–136.
Smith, T.C. 1961. Japan's aristocratic revolution. *Yale Review* 50(3): 370–383.
Spraos, J. 1982. Deteriorating terms of trade and beyond. *Trade and development. An UNCTAD review*, No. 4. Paris: UNCTAD.
Sutcliffe, R.B. 1971. *Industry and underdevelopment*. London: Addison-Wesley.
UNCTAD. 1981. *Fibres and textiles: Dimensions of corporate marketing structure*. Geneva: United Nations.
Von Tunzelmann, G.N. 1978. *Steam power and British industrialization to 1860*. Oxford: Clarendon Press.
Yamazawa, I., K. Taniguchi, and A. Hirata. 1983. Trade and industrial adjustment in Pacific Asian countries. *Developing Economies* 21(4): 281–312.

Inequalities

Peter Newman

Mathematical inequalities are pervasive in economic theory, just as economic inequalities are pervasive in social life. The insistence that quantities (always) and prices (usually) be nonnegative, the constraint that expenditure not exceed wealth, the necessity in proving existence of competitive equilibrium that each agent's resources have positive value, are so familiar that we

scarcely think of them as requirements of inequality, though that is what they are.

Many of the basic results of economic theory (such as the non-positivity of the substitution effect) take the form of inequalities. These in turn often arise from the definiteness or semi-definitenes of certain matrices, such definiteness being again expressed by inequalities. Yet further along the chain of reasoning, those matrices usually derive such properties from their origin in the convexity or concavity of various functions. For real-valued functions, convexity is defined by *Jensen's Inequality* (1906): The function $f: X \subset R^n \to R$ is *convex* if

$$\forall x^1, x^2 \in X, \quad \forall \alpha \in [0,1] \; f\left(\alpha x^1 + (1-\alpha)x^2\right) \times \leq \alpha f\left(x^1\right) + (1-\alpha) f\left(x^2\right) \quad (1)$$

(A function g is concave if $-g$ is convex).

There are close connections between convex functions and inequalities in general. Indeed, 'The classical inequalities are . . . obtained by verifying that a certain function is convex and by calculating its transforms.' (Young 1969, p. 112). To illustrate this general proposition by an important special case, consider the *gauge* $J(\cdot|C)$ of any set $C \subset R^n$, together with its *polar transform* $J^0(\cdot|C)$, which is the gauge of the polar set C^0 of C (see GAUGE FUNCTIONS). When C is convex and closed and contains the origin, $J^0(\cdot|C)$ becomes the *support function* $S(\cdot|C)$ of C. A fundamental inequality of convexity for gauges and their polar transforms is *Mahler's Inequality* (1939), which applied to the present situation reads:

$$\forall x \in R^n, \quad \forall y \in R^n \quad \sum x_i y_i \leq J(x|C) S(y|C) \quad (2)$$

Consider now R^n with its standard Euclidean norm $\| x\|_2 = \left(\sum x_i^2\right)^{1/2}$, and suppose that C is the closed unit sphere $Sc = \{x \in R^n : \| x\|_2 \leq 1\}$ of R^n. In this special case it happens that

$$J(\cdot|S_z) = \|\cdot\|_z = S(\cdot|S_c) \quad (3)$$

(see e.g. Rockafellar 1970, p. 130). So from (2) and (3),

$$\forall x \in R^n, \quad \forall y \in R^n \quad \sum x_i y_i = \left(\sum x_i^2\right)^{1/2} \quad \left(\sum y_i^2\right)^{1/2} \quad (4)$$

Since (4) is the famous *Cauchy–Buniakowski–Schwarz Inequality*, this illustrates Young's general proposition above. Young (1969, pp. 112–113) gives further examples (with proofs) of the connections between convexity and the classical inequalities, such as that relating the arithmetic and geometric means, and *Holder's Inequality* (1889):

$$\forall x \in R^n, \quad \forall y \in R^n \quad \sum x_i y_i \leq \left(\sum |x_i|^p\right)^{1/p} \quad \left(\sum |y_i|^q\right)^{1/q} \quad (5)$$

(where $p > 0$, $q > 0$, and $p^{-1} + q^{-1} = 1$), of which (4) is the special case $p = 2 = q$.

It is not surprising then that the classic work on inequalities, the delightful and indispensable book by Hardy et al. (1934), contains one of the earliest systematic treatments of convex functions in English. A later survey is Beckenbach and Bellman (1961).

See Also

- Convex Programming
- Gauge Functions

Bibliography

Beckenbach, E.F., and R. Bellman. 1961. *Inequalities*. Berlin: Springer.

Hardy, G.H., J.E. Littlewood, and G. Polya. 1934. *Inequalities*, 2nd ed. Cambridge: Cambridge University Press, 1952.

Hölder, O. 1889. Über einen Mittelwertsatz. *Göttinger Nachrichten*, 38–47.

Jensen, J.L.W.V. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* 30: 175–193.

Mahler, K. 1939. Ein Übertragungsprinzip für konvexe Körper. *Časopis Pêestováni Matematiky a Fysiky* 63: 93–102.

Rockafellar, R.T. 1970. *Convex analysis*. Princeton: Princeton University Press.
Young, L.C. 1969. *Lectures on the calculus of variations and optimal control theory*. Philadelphia: W.B. Saunders Company.

Inequality

Erik Olin Wright

To speak of a social inequality is to describe some valued attribute which can be distributed across the relevant units of a society in different quantities, where 'inequality' therefore implies that different units possess different amounts of this attribute. The units can be individuals, families, social groups, communities, nations; the attributes include such things as income, wealth, status, knowledge, power. The study of inequality then consists of explaining the determinants and consequences of the distribution of these attributes across the appropriate units.

This essay on Inequality has four principal objectives. First, I will propose a general typology of *forms of inequality*. This typology will help to map out the conceptual terrain of the discussion. Second, I will examine debates on the conceptual status of one particular type of inequality within this typology, inequality in material welfare. In particular, I will examine the debate over whether or not material inequalities in contemporary societies should be viewed as rooted in *exploitation*. Third, I will examine the implications of these contending views of material inequality for strategies for empirical research on income inequality. Finally, I will discuss the relationship between contending accounts of income inequality and the analysis of social classes.

A Typology of Inequalities

Social inequalities can be distinguished along two dimensions: first, whether the unequally distributed attribute in question is a *monadic* attribute or a *relational* attribute; and second, whether the process of acquisition of a particular magnitude of this attribute by the individual can be considered a monadic or relational *process*.

Monadic and Relational Attributes

A monadic attribute is any property of a given unit (individual, family, community, etc.) whose magnitude can be defined without any reference to other units. Material consumption is a good example: one can assess how much an individual unit consumes in either real terms or monetary terms without knowing how much any other unit consumes. This does not mean that the attribute in question has no social content to it. Monetary income, for example, is certainly a social category: having an annual income of a $30,000 only represents a source of inequality given that other people are willing to exchange commodities for that income, and this implies that the income has an irreducibly social content to it. Nevertheless, income is a monadic attribute in the present sense in so far as one can measure its magnitude without knowing the income of other units. Of course, we would not know whether this magnitude was high or low – that requires comparisons with other units. But the magnitude of any given unit is measurable independently of any other unit.

Relational attributes, in contrast, cannot be defined independently of other units. 'Power' is a good example. As Jon Elster (1985, p. 94) writes, 'In one simple conceptualization of power, my amount of power is defined by the number of people *over whom* I have control, so the relational character of power appears explicitly.' To be powerless is to be controlled by others; to be powerful is to control others. It is impossible to measure the power of any unit without reference to the power of others.

Monadic and Relational Processes

Certain unequally distributed attributes are acquired through what can be called a monadic process. To describe the distribution process (as opposed to the attribute itself) as monadic is to say that the immediate mechanisms which cause the magnitude in question are attached to

the individual units and generate their effects autonomously from other units.

A simple example of a monadic process that generates inequalities is the distribution of body weight in a population. The distribution of weight in a population of adults is certainly unequal – some people weigh three times the average weight of the population, some people weigh half as much as the average. An individual's weight is a monadic attribute – it can be measured independently of the weight of any other individual. And the weight acquisition process is also essentially monadic: it is the result of mechanisms (genes, eating habits, etc.) directly attached to the individual. This is not to say, of course, that these mechanisms are not themselves shaped by social (relational) causes: social causes may influence genetic endowments (through marriage patterns – e.g. norms governing skinny people marrying fat people) and social causes may shape eating habits. Such social explanations of body weight distributions, however, would still generally be part of a monadic process in the following sense: social causes may help to explain why individuals have the weight-regulating mechanisms they have (genes, habits), but the actual weight of any given individual results from these individual weight-regulating mechanisms acting in isolation from the weight-regulating mechanisms of other individuals. The empirical distribution of weights in the population is therefore simply the sum of these monadic processes of the individuals within the distribution.

Now, we can imagine a social process through which weight was determined in which this description would be radically unsatisfactory. Imagine a society in which there was insufficient food for every member of the society to be adequately nourished, and further, that social power among individuals determined how much food each individual consumed. Under these conditions there is a *causal* relation between how much food a fat (powerful) person eats and how little is consumed by a skinny (powerless) person. In such a situation, the immediate explanation of any given individual's consumption of food depends upon the social *relations* that link that individual to others, not simply on monadic mechanisms. Such an inequality generating process, therefore, would be described as relational rather than a monadic process. More generally, to describe the process by which inequalities are generated as relational, therefore, is to say that the mechanisms which determine the magnitude of the unequally distributed attribute for each individual unit causally depends upon the mechanisms generating the magnitude for other individuals.

Taking these two dimensions of inequality together, we can generate the following typology of ideal-typical forms of inequality. This typology (Table 1) is deliberately a simplification: the causal processes underlying the distribution of most inequalities will involve both monadic and relational mechanisms. Nevertheless, the simplification will help to clarify the conceptual map of inequalities which we have been discussing.

'Power' is perhaps the paradigmatic example of a relationally determined relational inequality. Not only is power measurable only relationally, but power is acquired and distributed through a relational process of competition and conflict between contending individuals, groups, nations, etc. (For discussions of power as form of inequality see Lenski 1966; Lukes 1974.)

Power is not, however, the only example. Social status is also generally an example of a relationally determined relational attribute. Status is intrinsically a relational attribute in that 'high' status only has meaning relative to lower statuses; there is no absolute metric of status. The process of acquisition of such high status is also generally a relational process of exclusion of rival contenders for status through competitive and coercive means. (Under special circumstances status-acquisition may be a largely monadic process. In

Inequality, Table 1 Typology of forms of inequality

		Form of the unequal attribute	
		Relational	Monadic
Form of the Process of Distribution of Attributes	Relational	Power, status Talent	Income Health, Weight
	Monadic		

artistic production, for example, one could imagine a situation in which each individual simply does the best he or she can and achieves a certain level of performance. There is nothing in one person's achievement of a given level of performance that precludes anyone else achieving a similar level. The status that results from that achievement, however, is still relational: if many people achieve the highest possible level of performance, then this level accords them less status than if few do, but the acquisition process would not itself be a relational one. In general, however, since the process by which the level of performance itself is achieved is a competitive one in which people are excluded from facilities for learning and enhancing performance, status acquisition is itself a relational process.)

The distribution of health is largely a monadic process for the distribution of a monadic attribute. In general, as in the weight acquisition case, the mechanisms which determine an individual's health – genetic dispositions, personal habits, etc. – do not causally affect the health of anyone else. There are, however, two important kinds of exceptions to this monadic causal process, both of which imply a relational process for the distribution of health as a monadic inequality. First, infectious diseases are clearly an example of a process through which the mechanisms affecting health in one person causally affect the health of another. More significantly for social theory, where the distribution of health in a population is shaped by the distribution of medical services, and medical services are relatively fixed in quantity and unequally distributed, then the causal mechanism producing health in one person may well affect the health of another in a relational manner.

Talent is an example of a relational attribute that is unequally distributed through a monadic process. A 'talent' can be viewed as a particular kind of genetic endowment – one that enhances the individual's ability to acquire various skills. To be musically talented means to be able to learn to play and compose music easily, not actually to play and compose music well (a potential prodigy who has never seen a piano cannot play it well). Talents are caused through a monadic process since the causal mechanism which determines one person's latent capacities to acquire skills does not affect anyone else's. (Obviously, parents' talent-generating mechanisms – genes – can affect their children's through inheritance. This is identical to the effect of parents' genes in the weight example. The point is that the effectiveness of one person's genes is independent of anyone else's.) The attribute so produced, however, is clearly relational: a talent is only a talent by virtue of being a deviation from the norm. If everyone had the same capacity to write music as Mozart, he would not have been considered talented.

Income inequality, at least according to certain theories of income determination (see below), could be viewed as an example of a relational process for distributing a monadic attribute. Income is a monadic attribute in so far as one individual's income is definable independently of the income of anyone else. But the process of acquisition of income is plausibly a relational one: the mechanisms by which one person acquires an income causally affects the income of others.

Inequalities in Material Welfare: Achievement Versus Exploitation

More than any other single kind of inequality, inequality of material welfare has been the object of study by social scientists. Broadly speaking, there are two distinct conceptualizations which have dominated the analysis of this kind of inequality in market societies. These I will call the achievement and exploitation perspectives.

Achievement Models

The achievement model of income determination fundamentally views income acquisition as a process of individuals acquiring income as a return for their own efforts, past and present. The paradigm case would be two farmers on adjacent plots of land: one works hard and conscientiously, the other is lazy and irresponsible. Assuming no externalities, at the end of a production cycle one has twice the income of the other. This is clearly a monadic process producing a distribution of a monadic outcome.

The story then continues: the conscientious farmer saves and reinvests part of the income earned during the first cycle and thus expands production; the lazy farmer does not have anything left over to invest and thus continues production at the same level. The result is that over time the inequalities between the two farmers increases, but still through a strictly monadic process.

Eventually, because of a continually expanding scale of production, the conscientious farmer is unable to farm his/her entire assets through his/her own work. Meanwhile the lazy farmer has wasted his/her resources and is unable to support him/herself adequately on his/her land. The lazy farmer therefore goes to work as a wage-earner for the conscientious farmer. Now, clearly, a relational mechanism enters the analysis, since the farm labourer acquires income in a wage paid by the farmer-employer. However, in the theory of wage-determination adopted in these kinds of models in which the labourer is paid exactly the marginal product of labour, this wage is exactly equivalent to the income the labourer would have received simply by producing the same commodities on his/her own account for the market. The relational mechanism, therefore, simply mirrors the initial monadic process.

In such achievement models of income acquisition genuinely relational processes may exist, but generally speaking these have the conceptual status of deviations from the pure model reflecting various kinds of disequilibria. In the sociological versions of achievement models – typically referred to as 'status attainment' models of stratification – these deviations are treated as effects of various kinds of ascriptive factors (race, sex, ethnicity) which act as obstacles to 'equal opportunity'. (The best example of status attainment models of inequality is Sewell and Hauser 1975.) Similarly, in the economic versions of such models – generally referred to as 'human capital' models – the deviations either reflect transitory market disequilibria or the effects of various kinds of extra-economic discrimination. (The classic account of human capital theory is given by Becker 1975. For his analysis of discrimination see Becker 1971.) In both the sociological and economic versions, these relational mechanisms of income determination that produce deviations from the pure achievement models mean that certain kinds of people are prevented from getting full income pay-offs from their individual efforts. The inner logic of the process, in short, is monadic with contingent relational disturbances.

Exploitation Models

Exploitation models of income inequality regard the income distribution process as fundamentally relational. The basic argument is as follows: In order to obtain income, people enter into a variety of different kinds of social relations. These will vary historically and can be broadly classified as based in different 'modes of production'. Through a variety of different mechanisms, these relations enable one group of people to appropriate the fruits of labour of another group (Cohen 1979). This appropriation is called exploitation. Exploitation implies that the income of the exploiting group at least in part depends on the efforts of the exploited group rather than simply their own effort. It is in this sense that income inequality generated within exploitative modes of production is intrinsically relational.

There are a variety of different concepts of exploitation contending in current debates. The most promising, in my judgement, is based on the work of Roemer (1983). (For a debate over Roemer's formulation, see *Politics & Society*, 11(2), 1982.) In Roemer's account, different forms of exploitation are rooted in different forms of property relations, based on the ownership of different kinds of productive assets. Roemer emphasizes two types of property in his analysis: property in the means of production (or alienable assets) and property in skills (or inalienable assets). Unequal distribution of the first of these constitutes the basis for capitalist exploitation; unequal distribution of the second constitutes the basis, in his analysis, for socialist exploitation.

While Roemer criticizes the labour theory of value as a technical basis for analysing capitalist exploitation, nevertheless his basic defence of the logic of capitalist exploitation is quite in tune with traditional Marxist intuitions: capitalists

appropriate part of the surplus produced by workers by virtue of having exclusive ownership of the means of production. Socialist or skill exploitation is a less familiar notion. Such exploitation is reflected in income returns to skills which is out of proportion to the costs of acquiring the skills. Typically this disproportion – or 'rent' component of the wage – will be reproduced through the institutionalization of credentials. Credentials, therefore, constitute the legal form of property that typically underwrites exploitation based in skills.

Two additional assets can be added to Roemer's analysis. Unequal distribution of *labour power* assets can be seen as the basis for feudal exploitation, and unequal distribution of *organization* assets can be viewed as the basis for state bureaucratic exploitation (i.e. the distinctive form of exploitation in 'actually existing socialism'). The argument for feudalism is basically as follows: in feudal society, individual serfs own less than one unit of labour power (i.e. they do not fully own their own labour power) while the lord owns part of the labour power of each of his serfs. The property right in the serf's labour power is the basis for the lord forcing the serf to work on the manorial land in the case of corv*é*e labour, or paying feudal rents in cases where corv*é*e labour has been converted into other forms of payment. The flight of peasants to the cities, in these terms, is a form of theft from the lord: the theft of the lord's labour power assets. The argument for state bureaucratic societies is based on the claim that control over the organizational resources of production – basically control over the planning and coordination of the division of labour – is the material basis for appropriation of the surplus by state bureaucrats. (For a detailed discussion of these additional types of assets and their relationship to exploitation, see Wright 1985.) In all of these cases, the ownership and/or control of particular types of productive assets enables one class to appropriate part of the social surplus produced by other classes.

In exploitation models of income distribution, monadic processes can have some effects. Some income differences, for example, may simply reflect different preferences of individuals for work and leisure (or other trade-offs). Some of the income difference across skills may simply reflect different costs of acquiring the skills and therefore have nothing to do with exploitation. Such monadic process of income determination, however, are secondary to the more fundamental relational mechanisms.

Implications for Empirical Research Strategies

As one would suspect, rather different empirical research strategies follow from monadic versus relational conceptions of the process of generating income inequality. In a strictly monadic approach, a full account of the individual (non-relational) determinants of individual income is sufficient to explain the overall distribution of income. This suggests that the central empirical task is first, to assemble an inventory of all of the individual attributes that influence the income of individuals, and second, to evaluate their relative contributions to explaining variance across individuals in income attainment. In the case of the example of the two farmers discussed above this would mean examining the relative influence of family background, personalities, education and other individual attributes in accounting for their different performances. The sum of such explanations of autonomously determined individual outcomes would constitute the basic explanation of the aggregate income distribution.

It follows from this that the heart of statistical studies of income inequality within an achievement perspective would be multivariate micro-analyses of variations in income across individuals. The study of overall income distributions as such would have a strictly secondary role.

In exploitation models of income distribution, the central empirical problem is to investigate the relationship between the variability in the form and degree of exploitation and income inequality. This implies a variety of specific research tasks, including such things as studying the relationship between the overall distribution of exploitation-generating assets in a society and its overall distribution of income, the different processes of

income determination within different relationally defined class positions (see Wright 1979), and the effects of various forms of collective struggle which potentially can counteract (or intensify) the effects of exploitation-mechanisms on income inequalities.

This does not imply, of course, that achievement models of income inequality have no interest in macro-studies of income distribution, nor that exploitation models have no interest in micro-studies of individual income determination. But it does mean that the core empirical agendas of each model of income inequality will generally be quite different.

Material Inequality and Class Analysis

Sociologists are interested in inequalities of material welfare not simply for their own sake, but because such inequality is thought to be consequential for various other social phenomena. Above all, material inequality is one of the central factors underlying the formation of social classes and class conflict.

The two models of income inequality we have been discussing have radically different implications for class analysis. In achievement models of income distribution, there is nothing intrinsically antagonistic about the interests implicated in the income determination process. In the example we discussed, the material interests of the lazy farmer are in no sense intrinsically opposed to those of the industrious farmer. The strictly economic logic of the system, therefore, generates autonomous interests of different economic actors, not conflictual ones.

Contingently, of course, there may be conflicts of interest in the income determination process. This is particularly the case where discrimination of various sorts creates noncompetitive privileges based on ascriptive characteristics such as sex and race. These conflicts, however, are not fundamental to the logic of market economies and they do not constitute the basis for conflicts between economic classes as such.

Conflicts between classes in capitalist societies, therefore, basically reflect either cognitive distortions on the part of economic actors (e.g. misperceptions of the causes of inequality) or irrational motivations (e.g. envy). Conflicts do not grow out of any objective antagonism of interests rooted in the very relations through which income inequalities are generated.

Exploitation models of income inequality, in contrast, see class conflict as structured by the inherently antagonistic logic of the relational process of income determination. Workers and capitalists have fundamentally opposed interests in so far as the income of capitalists depends upon the exploitation of workers. Conflict, therefore, is not a contingent fact of particular market situations, nor does it reflect ideological mystifications of economic actors; conflict is organic to the structure of the inequality-generating mechanisms themselves.

These different stances towards the relationship between interests and inequality in the two approaches means that for each perspective different social facts are treated as theoretically problematic, requiring special explanations: conflict for achievement theories, consensus for exploitation theories. Both models, however, tend to explain their respective problematic facts through the same kinds of factors, namely combinations of ideology and deviations from the pure logic of the competitive market. Exploitation theories typically explain cooperation between antagonistic class actors on the basis 'false consciousness' and various types of 'class compromises' between capitalists and workers, typically institutionalized through the state, which modify the operation of the market (see Przeworski 1985). Achievement theories, on the other hand, use discriminatory preferences and market imperfections to explain conflict.

See Also

- ▶ Capital as a Social Relation
- ▶ Class
- ▶ Distributive Justice
- ▶ Economic Freedom
- ▶ Equality
- ▶ Justice

- ▶ Poverty
- ▶ Property

Bibliography

Becker, G.S. 1971. *The economics of discrimination*, 2nd ed. Chicago: University of Chicago Press.
Becker, G.S. 1975. *Human capital*, 2nd ed. New York: National Bureau of Economic Research.
Cohen, G.A. 1979. The labor theory of value and the concept of exploitation. *Philosophy and Public Affairs* 8(4): 338–360.
Elster, J. 1985. *Making sense of Marx*. Cambridge: Cambridge University Press.
Lenski, G. 1966. *Power and privilege*. New York: McGraw-Hill. *Politics & society*. 1982. 11(3). Special issue on John Roemer's theory of class and exploitation.
Lukes, S. 1974. *Power: A radical view*. London: Macmillan.
Przeworski, A. 1985. *Capitalism and social democracy*. Cambridge: Cambridge University Press.
Reich, M. 1981. *Racial inequality*. Princeton: Princeton University Press.
Roemer, J. 1983. *A general theory of exploitation and class*. Cambridge, MA: Harvard University Press.
Sewell, W., and R. Hauser. 1975. *Education, occupation and earnings*. New York: Academic.
Szymanski, A. 1976. Racial discrimination and white gain. *American Sociological Review* 41: 403–414.
Wright, E.O. 1979. *Class structure and income determination*. New York: Academic.
Wright, E.O. 1985. *Classes*. London: New Left Books/ Verso.

Inequality (Global)

Steve Dowrick

Abstract

The seemingly inexorable rise in global inequality from the early 19th century may have reached a plateau at the end of the 20th century, although there are disputes about the methodology underlying that conclusion. Increasing global inequality in the 20th century was driven largely by increasing income gaps between nations. Inequality within countries fell sharply at the beginning of the 20th century, rising slightly towards the end. The strong economic growth of the Chinese economy is tending to reduce global inequality as China moves up towards the middle of the income ladder.

Keywords

Balassa–Samuelson effect; Catch-up; Geary–Khamis purchasing power parity; Gini index; Inequality (global); Inequality (measurement); Inter-country inequality; Intracountry inequality; Kuznets, S.; Mean logarithmic deviation; National accounting; Purchasing power parity; Standard deviation of logarithmic income; Theil index; Uneven development

JEL Classifications

O1; E3

The seemingly inexorable rise in global inequality in the 19th and 20th centuries may have reached a plateau in the 1980s.

The causes and consequences of changing global inequality are a hotly contested area of economic research and debate. The intensity of the debate is in part due to the moral outrage felt by many at revelations such as those from the International Comparison Program (2007), henceforth ICP. The ICP 1996 data on average real expenditures per person reveal expenditure exceeding 1,000 dollars on the luxuries of alcoholic beverages, recreation and restaurant meals in each of the world's 20 richest countries, an amount that exceeds the total national income in each of the world's 12 poorest countries and exceeds total expenditure on food in each of the world's 70 poorest countries.

Income distribution estimates reveal that in the year 2000 more than one in ten of the world's population eked out a living around or below the World Bank's intermediate poverty line of two dollars per person per day, whilst the richest five per cent enjoyed incomes at or above 100 dollars per person per day. According to the World Bank (2006a), out of every 100 child born today, less

than one child in the USA is expected to die before the age of five, but for children born in Mali, 24 children will not survive.

The extent of current global inequality far exceeds the inequalities of previous eras, apparently giving the lie to theories that the forces of global integration reduce inequality through factor-price equalizing trade, boosting demand for low-wage labour in the poorest countries, and through capital mobility, whereby global investment flows to the poorest and least capital-intensive countries, boosting labour productivity and real wages – although these observations must be tempered by the evidence that some aggregate measures of global inequality peaked towards the end of the 20th century and by the evidence of the highly successful catch-up growth of many East Asian economies in the second half of that century.

There are many problems in conceptualizing and measuring inequality: are we concerned with measured incomes, with consumption or with well-being? Is inequality measured across nations, across households or across individuals? What is the appropriate index of inequality to use? For the most part I will focus on inequalities in measured income based on national accounting conventions or on survey data. Rather than debate the merits of different indices of inequality, I report a range of commonly used measures – noting that many studies find that different indices tend to move in the same direction over time even if their levels differ. Towards the end of this article I consider some of the methodological problems.

Inequality Over the Centuries

Looking back to the year 1500, Angus Maddison (2003) has dared to publish estimates of average income levels – or, more precisely, real GDP per capita measured at 1990 international prices, which I refer to as 'income' for short. His estimates suggest that over the first three centuries global income rose very slowly – from 566 dollars per person in 1500 to 667 dollars in 1820. Over this period, national income levels did not differ by very much, most of the nations being less than 50 per cent above or below the world average. As world income growth began to accelerate through the 19th and 20th centuries, led first by the United Kingdom and then by the United States, income gaps began to widen. By the end of the 20th century the world's richest major nation, the United States, was more than 100 times richer than the world's poorest nation.

These broad trends in growth and inequality are illustrated in Fig. 1, which displays average income levels across eight populous countries and regions at approximately 50-year intervals from 1500 to 2000. Averaging incomes across regions does of course understate the true extent of inter-country inequality, particularly in the case of Africa where the 2000 average of nearly 1,500 dollars disguises a maximum income of over 10,000 dollars in Mauritius and a minimum of just 218 dollars in Zaire.

It is also the case that averaging incomes within countries disguises the true extent of inequality across individuals or households (or inequality by gender or ethnic groups). The paucity of historical data on income distribution within countries makes disaggregation below the national level an extremely difficult task for eras before the late 20th century. This task has, however, been attempted by François Bourguignon and Christian Morrisson (2002), who estimate global inequality across a group of 33 countries/country-groups reaching back to 1820 using historical income distribution data and extrapolating across countries judged to be similar. Their results are displayed as the four solid lines in Fig. 2.

It is apparent that global income inequality rose strongly in the 19th century on all four of their measures: the Gini index, the Theil index, the mean logarithmic deviation (MLD) and the standard deviation of logarithmic income. Bourguignon and Morrisson's estimates indicate a slowing down in the rate of increase in inequality in the 20th century, although each measure displays slightly different trends. The Gini flattens out after 1970, both of the logarithmic measures peak in 1980, whilst the Theil measure is flat between 1910 and 1970 but rises up to 1992.

Both the Theil and the MLD can be decomposed exactly into the contributions of

inequality within countries and inequality between countries. The within-country contributions to global inequality are shown as the dashed lines in Fig. 2. It is apparent that within-country inequality was high and stable in the 19th century but fell substantially in the first half of the 20th century. On both measures, the contribution of within-country inequality to total inequality fell from nearly 90 per cent in 1820 to 40 per cent over the second half of the 20th century.

Global Income Inequality in the Late 20th Century

Data availability is far less of a problem for the second half of the 20th century than for previous eras (though problems of data definition and reliability persist) due to the publication of time series data on real GDP across most of the world's economies by Maddison (2003), by Robert Summers and Alan Heston (1991) and by Heston, Summers and Bettina Aten (2002) – the latter two studies producing successive versions of the Penn World Table. All these authors extrapolate over time and across countries from the benchmark price surveys, which are carried out periodically by the International Comparison Program.

Klaus Deininger and Lyn Squire (1996) have compiled sporadic time series on income distribution within countries – typically by decile or quintile groups. The gaps in their annual and country coverage have been filled by James K. Galbraith and Hyunsub Kum (2003), who extrapolate using data on wage inequality. Branko Milanovic (2002, 2005) has independently compiled a large number of national surveys of the distribution of income or expenditure at household level. I draw on a number of studies that have analysed global inequality using these sources of data.

A majority of these studies concludes that global income inequality peaked in the 1970s or 1980s and has subsequently declined slightly. The majority position has been challenged by Milanovic (2002, 2005) who uses household income surveys and World Bank estimates of current purchasing power of currencies to show that global inequality rose in the 1990s, in contradiction to Xavier Sala-i-Martin (2006) who demonstrates falling global inequality over the same period. Sala-i-Martin's methodology differs from that of Milanovic in that he uses the Deininger and Squire data on within-country inequality and converts currencies using the constant price estimates of purchasing power parity from the Penn World Table. The majority position is also contested by Steve Dowrick and Mohammed Akmal (2005), who show that the Penn World Table's method of measuring real GDP at constant prices is subject to time-varying substitution bias, which understates the true level of inequality across countries. The evidence on this debate from Bourguignon and Morrisson (2002) is equivocal since two of their measures of global inequality, the standard deviation and the mean deviation of logarithmic income, fall between 1980 and 2000 whilst their other two measures, the Gini and Theil indices, are flat or rising after 1980.

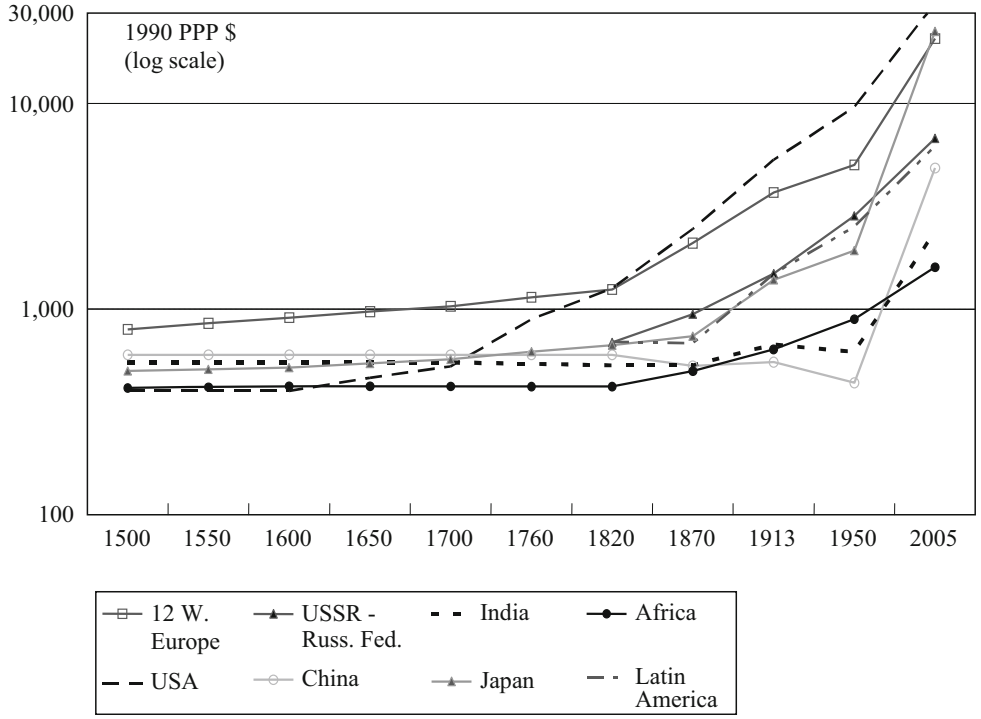
This debate on recent trends is important in that it identifies key methodological problems and it emphasizes the fact that any attempt to measure global inequality is subject to a considerable margin of error. The debate is heated because the majority view can be interpreted as support for the equalizing tendencies of global capitalism, giving some comfort to those embarrassed by the evidence of relentless growth in inequality.

Nevertheless, the 'big pictures' of both Maddison (2003) and Bourguignon and Morrisson (2002) – see Figs. 1 and 2 – prevail. After 150 years of unparalleled growth and rising inequality, global inequality appears to have stabilized towards the end of the 20th century.

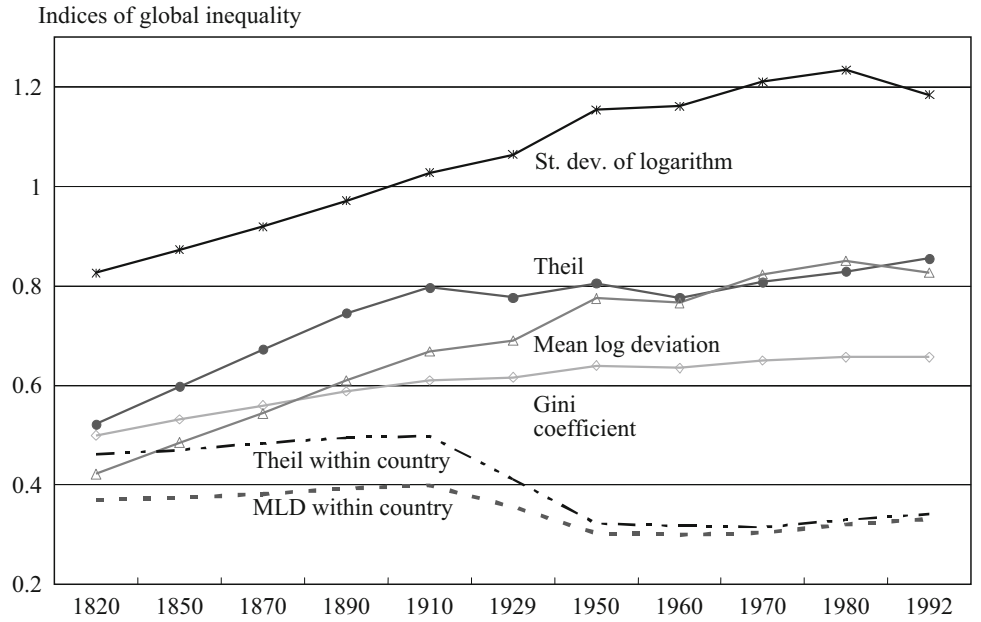
Decomposing Global Income Inequality

Within the context of this big picture, I will examine the principal components that contribute to the overall extent of global inequality: inequality across countries; weighting countries by population; and inequalities within countries.

Examining inequality in national average incomes (or GDP per capita) has been part of the focus of research into economic growth and



Inequality (Global), Fig. 1 Long-run development: real GDP per capita, 1500–2005 (*Note*: Data for 1550, 1650 and 1760 have been interpolated. *Source*: Maddison (2003) extended to 2005 using World Bank (2006b))

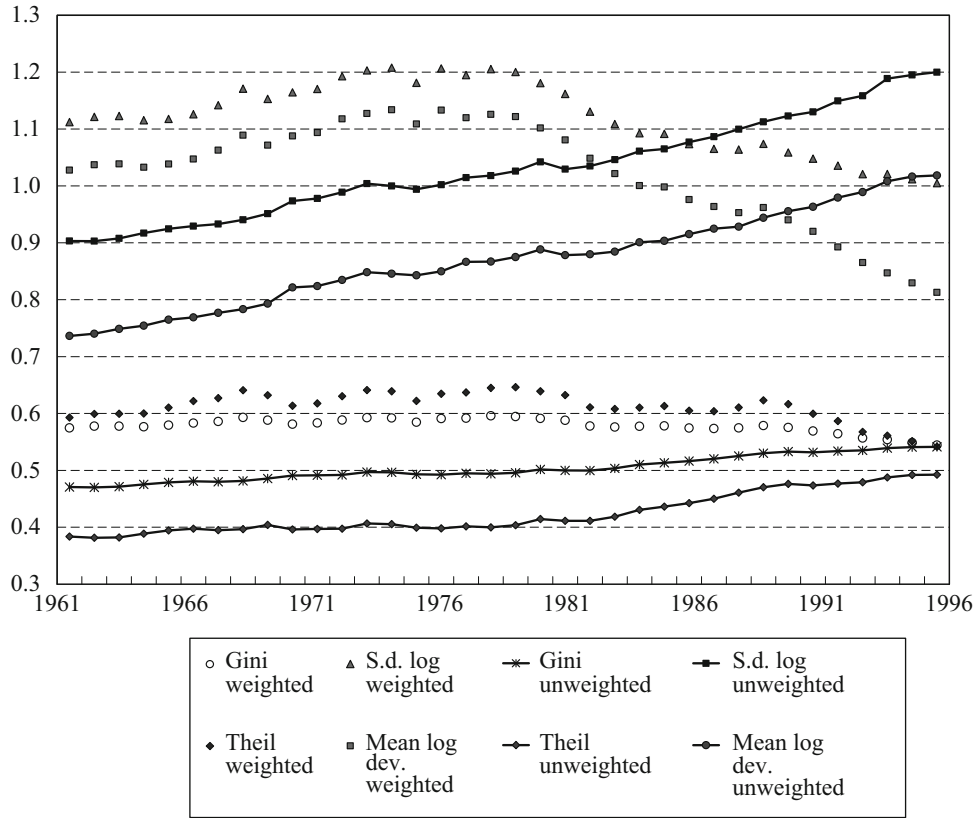


Inequality (Global), Fig. 2 Inequality within and between 33 countries, 1820–1992 (*Source*: Bourguignon and Morrison (2002))

convergence. The consensus in that literature has been aptly summarized in the title of a paper by Lant Pritchett (1997), 'Divergence, big time'. Some of the growth research has concentrated on evidence of conditional convergence, whereby there is a tendency for poorer countries to grow faster than richer countries provided that some growth determinants are held constant. Conditional convergence is not, however, a sufficient condition for inequality to fall over time, since random shocks will tend to increase dispersion of income levels, and many of the common conditioning factors, such as investment rates or levels of human capital, are distributed in such a way as to limit the growth rates of the poorer countries. So there is no logical contradiction between evidence of conditional convergence and evidence of increasing inequality between countries.

Trends in inter-country inequality are illustrated in Fig. 3, where I plot four measures of inequality across 112 countries which together account for nearly 90 per cent of the global population. The time series are represented by the four solid lines. All four measures trend upwards between 1961 and 1996.

The proximate causes of this rise in inequality between countries in the period include the relatively rapid growth of the already rich United States (averaging 2.2 per cent per year in growth of real GDP per capita), the even faster growth of the relatively rich economies of western Europe (averaging 2.7 per cent per year) which have benefited from technological catch-up with the USA, and the tragedy of African economies which, on average, recorded less than one per cent growth per year. Fifteen African economies experienced falling income levels. With the rich



Inequality (Global), Fig. 3 Inequality across 112 countries, 1961–1996: population weighted and unweighted (*Source*: Penn World Table 6.1 (Heston et al. 2002))

nations becoming relatively richer and the poorest nations becoming relatively poorer, it is no surprise that all four measures of inter-country inequality record increases.

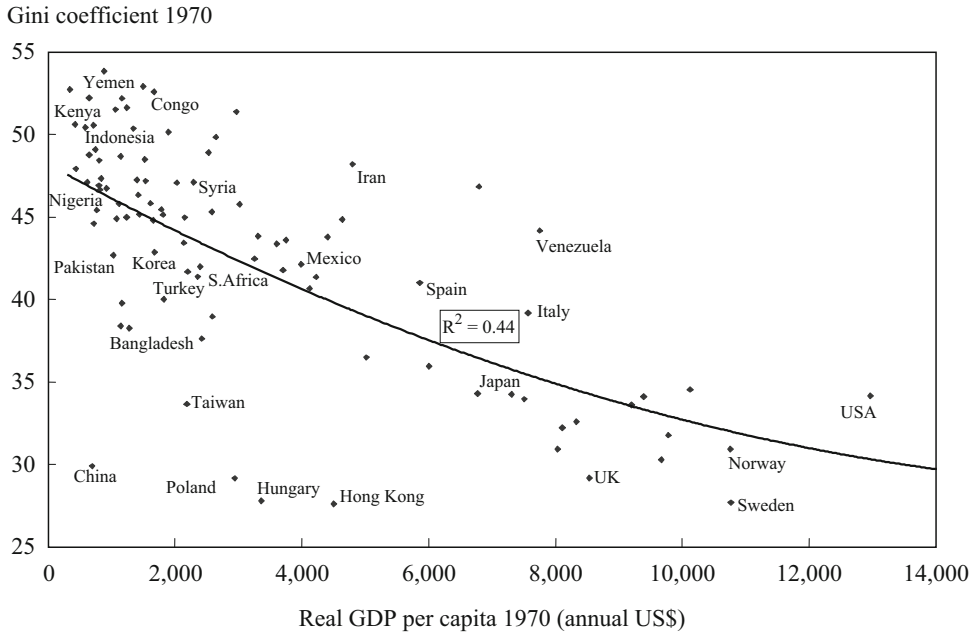
These comparisons, in the tradition of the literature on economic growth, give equal weight to each country. When examining inequality, however, we are often interested in inequality across households or individuals, so it makes sense to weight each country's average income by the population of that country. As many researchers have pointed out, this procedure changes the picture drastically – as illustrated by the dashed lines in Fig. 3, which are non-monotonic.

All four of the population-weighted measures of inequality between countries reach a peak in the late 1970s. This peak corresponds to the time when the growth rate of the Chinese economy took off. Through the 1960s and 1970s the Chinese economy grew at a moderate rate, moving average income from 21 per cent of the world average in 1960 to 26 per cent by 1978, still below African income levels. Over the next two decades, the growth rate accelerated, moving Chinese average income in 1996 up to 69 per cent of the world average. This movement of one-fifth of the world's population away from the bottom and towards the middle of the country income distribution is the principal cause of the substantial fall in population-weighted inequality across countries. Another contributory factor was the rise in the growth rate of the Indian economy, which moved from income at 21 per cent of the world average in 1980 to 32 per cent by 2000. (Relative income levels are derived from Maddison 2003.)

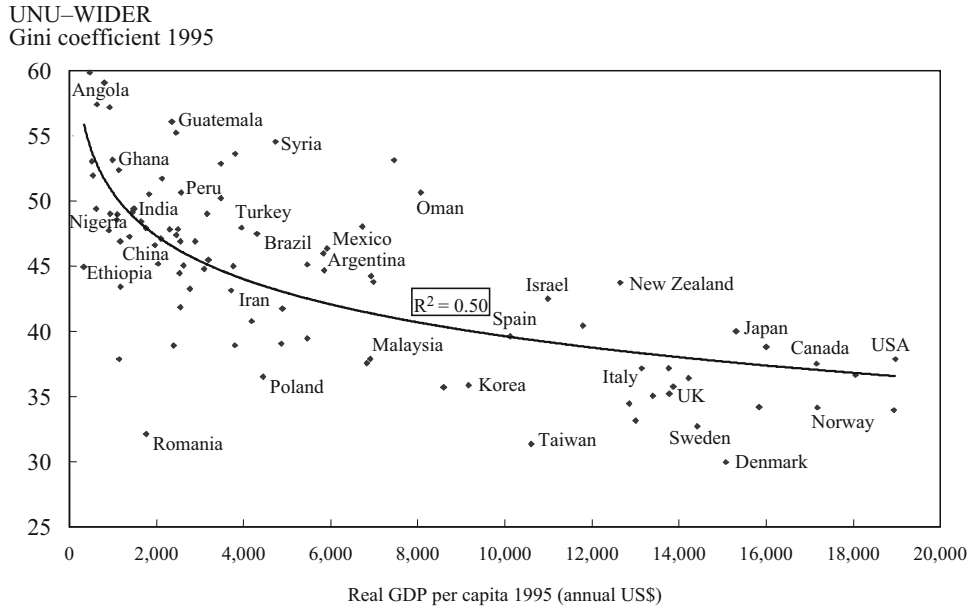
The final dimension to global inequality is inequality within countries. There has been widespread concern within the rich industrialized economies that the rapid expansion in the 1980s and 1990s of trade with low-wage economies such as China would cause increasing inequality as less skilled workers faced wage cuts or unemployment in the face of competitive imports. At the same time, real wages were rising for workers in developing economies who found jobs in the expanding export sectors. Indeed, it has been the case that many of the richest economies have experienced rising income inequality, with Gini coefficients averaging a rise of 3.5 points between 1970 and 1995 in the richer half of the sample of countries. Income inequality also increased in many of the poorer countries, averaging a rise of 2.2 points. (Data on inequality within countries are from Galbraith and Kum 2003, supplemented by estimates for China in 1970 and 1995 from the UNU-WIIDER data-set, sourced to Dowling and Soo 1983, and to Khan and Riskin 1998, respectively.)

Kuznets (1955) has famously observed that, over the course of economic development in the 19th century and the first half of the 20th century, income inequality first rose as labour moved from agriculture into industrial sectors with higher wages and then declined as industrial employment stabilized and wages were equalized. The ensuing implication of a hump-shaped cross-sectional relationship between inequality and income levels is not, however, supported by the crosssectional evidence from 1970 and 1995, which is illustrated in Figs. 4 and 5. Each figure plots the Gini coefficient on the vertical axis against the income level. The best-fit quadratic regression line has been added to each figure. For each year, it is evident that there is a fairly strong tendency for income inequality to fall as average income levels rise. This graphical analysis confirms the results of the econometric study conducted by Schultz (1998).

Comparison of Figs. 4 and 5 confirms the tendency for inequality to have risen within countries over the 25-year period. It is of particular interest to note the sharp rise in estimated inequality for China, from a Gini around 30 in 1970 – commensurate with the low levels of inequality observed in the Communist countries of eastern Europe – to a Gini of 45 in 1995 – commensurate with the more generally observed levels of inequality amongst other countries at the same level of development. This sharp rise in inequality is in keeping with accounts of rising inequality between the provinces in China, reflecting uneven development between rural regions and the rapidly industrializing coastal cities. Over the same period, the Indian Gini



Inequality (Global), Fig. 4 Intra-country inequality, 1970 (*Source*: Galbraith and Kum (2003) and UNU–WIDER (2005))



Inequality (Global), Fig. 5 Intra-country inequality, 1995 (*Source*: Galbraith and Kum (2003) and UNU–WIDER (2005))

coefficient of inequality was fairly stable at 46.9 in 1970 and 47.2 in 1995.

It might be expected that the general rise in inequality within countries after 1970, particularly within China, would offset any tendency for population-weighted inequality between countries to decline in the 1980s and 1990s. There is some evidence of this offsetting in the Bourguignon and Morrisson (2002, Table 2) data on global inequality, which is illustrated in Fig. 2. Two of their measures, the Theil index and the MLD, allow an exact decomposition into within-country and between-country inequality. The within component of the Theil index rises from 0.315 in 1970 to 0.342 in 1992, whilst the within component of the MLD rises from 0.304 to 0.332. Their Theil measure of global inequality does indeed continue to rise after 1980, although the MLD falls slightly.

Similar results on within-country inequality are reported by Sala-i-Martin (2006), who finds that the within component of the Theil index rises from 0.255 to 0.284 between 1970 and 2000, and the within component of the MLD rises from 0.246 to 0.319. However, his methodology differs from Bourguignon and Morrisson (2002) in that he studies a much larger number of countries and uses nonparametric estimates of within-country distribution. His overall conclusion is that global inequality fell towards the end of the 20th century despite the rise in inequality within countries.

It is noteworthy that 20th century movements in within-country inequality tend to be dominated by movements in population-weighted inequality across countries. This is not surprising since the most glaring inequalities are found in comparisons across countries. Typical values for the quintile ratio (the income of the richest fifth relative to the income of the poorest fifth) are around seven or eight when we look within countries, but across countries the quintile ratio is over 20. This is very different from the situation at the beginning of the 19th century when the dominating influence on global inequality was the extent of inequality within countries.

Methods of Comparing Income Levels Across Countries

Most studies that examine population-weighted inequality between countries conclude that inequality peaked in the 1970s and declined in the 1980s and 1990s. These studies depend on estimates of GDP per capita evaluated at purchasing power parities (PPP) using data from either Maddison (2003) or the Penn World Table. Maddison's data is used by Bourguignon and Morrisson (2002) and by Sutcliffe (2004). The Penn World Table data are used by Schultz (1998), Firebaugh (1999), Melchior et al. (2000) and Sala-i-Martin (2006), among others.

Several of these studies have contrasted their results with those obtained by Korzeniewicz and Moran (1997) and the United Nations Development Report (UNDP 2006), who use market rates of exchange rather than PPP exchange rates to compare incomes across countries. The use of market exchange rates leads to the conclusion that income inequality across countries was rising rather than falling over the final decades of the 20th century. There is widespread agreement that exchange rate comparisons are not appropriate if income inequality measures are being calculated in an attempt to evaluate inequality in human welfare. They suffer from two major defects. Market exchange rates are volatile, implying unrealistically sharp short-term movements in real incomes. They also systematically exaggerate real income differentials due to the Balassa–Samuelson effect whereby market exchange rates take no account of the relative cheapness of non-traded goods and services in low-wage low-income countries. Market rates of exchange systematically undervalue incomes in poor countries.

There are, however, some purposes for which the exchange rate measures of inequality may be more appropriate than PPPs. If we are concerned with the ability of poor countries to catch up with the technologies of the rich, and if this depends on their ability to purchase high-tech equipment from the major exporters of capital equipment, then it is

the exchange rate which is the appropriate measure of their capacity to develop. The same may well be true when we consider the bargaining power of the poorer nations at international forums such as the World Trade Organization.

To the extent that we are interested in income comparison as an approximation to welfare comparison, it is clearly preferable to compare incomes across countries at purchasing power parity. There is, however, a complication: which measure of purchasing power parity should we use? The PPPs used by both Maddison and the Penn World Table rely on the Geary–Khamis method, which calculates a weighted average of relative prices across all of the countries surveyed by the International Comparison Project in a benchmark year and values the GDP bundles of all countries in all years at that fixed set of prices. The weighting procedure uses country expenditure shares in world GDP, generating 'world prices' which are close to the price relativities prevailing in the rich countries of the Organisation for Economic Co-operation and Development (OECD) but very different from the relative prices prevailing in the world's poorer economies. The Geary–Khamis procedure induces substitution bias, valuing the abundant and cheap local services in low-wage economies at the much higher relative price of the rich economies. The effect is the opposite of the bias in the exchange rate comparisons. The Geary–Khamis PPPs systematically overvalue incomes in poorer countries, resulting in measures of global income inequality which are biased downwards. Dowrick and Akmal (2005) demonstrate that the use of Geary–Khamis PPPs can also distort the trend, since the magnitude of the bias changes over time, and show that an unbiased measure of global inequality does not fall between 1980 and 1993.

Further problems with the standard methods of comparing incomes across countries are pointed out by Milanovic (2005). He argues that it is illogical or at least inconsistent to use household survey data to estimate income distribution within countries, but to use national accounting measures rather than the survey measures when computing differences in average income levels across countries. Using average survey income, converted at PPP, he finds that global inequality rose between 1988 and 1993 before falling slightly by 1998. Milanovic notes that average survey income is always less than national accounts measures of average income, or GDP per capita, because it omits public expenditures. Lacking data on the distribution of public expenditures, he argues that survey income is the preferable measure.

Concluding Remarks

Global income inequality rose to historically unprecedented proportions, morally repugnant to many, through most of the 19th and 20th centuries – at the same time as average income levels rose to previously unimaginable heights. Since the 1970s, the level of inequality appears to have halted or, by some measures, has begun to fall slightly.

Prospects for the future evolution of global inequality depend crucially on two questions. First, will China continue to follow the trail of development blazed by Japan and Korea several decades earlier? If one-fifth of the world's population does indeed follow this path, then we can expect measures of global inequality to fall as Chinese income level approach the world average; but inequality will then increase as Chinese income levels catch up with those of the global rich. Second, can the desperately poor nations of Africa find a way, with or without the assistance of the rest of the world, to follow the successful development path on which China and India embarked in the 1980s and the 1990s? If African development fails to take off and if population growth continues to exceed that of the other continents, then global inequality may well resume its rising trend in the course of the 21st century.

See Also

- ▶ Gini Ratio
- ▶ Inequality (Measurement)
- ▶ Kuznets, Simon (1901–1985)

Bibliography

Bourguignon, F., and C. Morrisson. 2002. Inequality among world citizens: 1820–1992. *American Economic Review* 92: 727–744.

Deininger, K., and L. Squire. 1996. A new data set measuring income inequality. *World Bank Economic Review* 10: 565–591.

Dowling, J.M., and D. Soo. 1983. *Income distribution and economic growth in developing Asian countries*. Staff Paper No. 15. Manila: Asian Development Bank.

Dowrick, S., and M. Akmal. 2005. Contradictory trends in global income inequality: A tale of two biases. *Review of Income and Wealth* 51: 201–229.

Firebaugh, G. 1999. Empirics of world income inequality. *American Journal of Sociology* 104: 1597–1630.

Galbraith, J.K., and H. Kum. 2003. *Estimating the inequality of household incomes: Filling gaps and fixing problems in Deininger and Squire*. Working paper, Inequality Project, University of Texas.

Heston, A., R. Summers, and B. Aten. 2002. Penn World Table Version 6.1. Center for International Comparisons, University of Pennsylvania.

International Comparison Program. 2007. 1996 price survey data: http://pwt.econ. upenn.edu/Downloads/benchmark/benchmark.html

Khan, A.R., and C. Riskin. 1998. Income and inequality in China: Composition, distribution and growth of household income. *China Quarterly* 154: 221–253.

Korzeniewicz, R.P., and T.P. Moran. 1997. World economic trends in the distribution of income, 1965–92. *American Journal of Sociology* 102: 1000–1039.

Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45: 1–28.

Maddison, A. 2003. *The world economy: Historical statistics*. Paris: OECD.

Melchior, A., K. Telle, and H. Wiig. 2000. *Globalization and inequality: World income distribution and living standards, 1960–1998*. Oslo: Norwegian Institute of Foreign Affairs.

Milanovic, B. 2002. True world income distribution, 1988 and 1993: First calculations based on household surveys alone. *Economic Journal* 112: 51–92.

Milanovic, B. 2005. *Worlds apart: Measuring international and global inequality*. Princeton and Oxford: Princeton University Press.

Pritchett, L. 1997. Divergence, big time. *Journal of Economic Perspectives* 11(3): 3–17.

Sala-i-Martin, X. 2006. The world distribution of income: Falling poverty and . . . convergence, period. *Quarterly Journal of Economics* 121: 351–397.

Schultz, T.P. 1998. Inequality in the distribution of personal income in the world: How it is changing and why. *Journal of Population Economics* 11: 307–344.

Summers, R., and A. Heston. 1991. The Penn World Table (Mark 5): An expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* 106: 327–368.

Sutcliffe, B. 2004. World inequality and globalization. *Oxford Review of Economic Policy* 20: 15–37.

UNDP (United Nations Development Programme). 2006. *Human Development Report 2006*. New York: Oxford University Press.

UNU–WIDER. 2005. World income inequality database 2.0a. Online. Available at http://www.wider.unu.edu/wiid/wiid.htm, Accessed 8 March 2007.

World Bank. 2006a. *World development report*. New York: Oxford University Press.

World Bank. 2006b. *World development indicators 2006*. Online. Available at http://go.worldbank.org/3SGLDH5V10, Accessed 12 April 2007.

Inequality (International Evidence)

Andrea Brandolini and Timothy M. Smeeding

I

Abstract

The methodological assumptions underlying international comparisons of levels and trends in inequality are discussed, starting with the choice of the evaluative space. Empirical evidence shows that at the end of the 1990s, the United States had the highest level of disposable income inequality among high-income economies, while northern and central European countries had the lowest levels. Only in Russia and Mexico, two middle-income economies, was disposable income more unequally distributed. No common trend in inequality is observed since the 1970s across rich nations. Public redistribution through taxes and benefits influence both levels and changes in inequality.

Keywords

Atkinson index; Capability approach; Consumer price index; Disposable income; Expenditure; Gini index; Human capital; Income; Income inequality; Inequality (measurement); Inequality, international evidence of; Kuznets, S.; Lorenz curve; Luxembourg Income Study; Market income; Pareto's law; Purchasing power parity; Redistribution of income; Relative inequality; Standard of living; Theil index

JEL Classifications
D31; H2; I3; D63; E25; D3

The comparison of inequality across countries and over time has a long tradition in economics. In 1897 Pareto used data from tax returns for a heterogeneous group of nations, spanning a period of almost four centuries, to conclude that income inequality was remarkably constant over time and space. An intense debate followed, such that the editors of *Econometrica* devoted the second 'Annual Survey of Statistical Data' to Pareto's law (Bresciani-Turroni 1939), which served to bring to an end the idea of a 'natural' constancy of the distribution of income.

The study of international differences in income distribution gathered new momentum after the Second World War. In the 1950s, United Nations agencies pioneered the assembly of international datasets on income inequality (for example, United Nations 1951) and Kuznets (1955) stated his celebrated hypothesis of an inverted-U relationship between inequality and growth. Since those early days, international agencies and individual scholars have increasingly been engaged in collecting information on income distribution and comparing levels and trends of inequality across nations (Gottschalk and Smeeding 1997; Atkinson and Brandolini 2001). Cross-country comparisons of income inequality have become common in analysis that informs policymaking: measures of income distribution are featured among the indicators of social cohesion agreed by the European Union to monitor the performance of member countries (Atkinson et al. 2002), and one of the first charts of the 2006 *World Development Report* ranks nations by the Gini index of income (or expenditure) to show that 'Africa and Latin America have the world's highest levels of inequality' (World Bank 2005, Fig. 2.9, p. 39; the underlying data are reported in Table 1).

Focal Variable

As Sen suggests, the relative advantages and disadvantages that people have, compared with each other, can be judged in terms of many different variables, e.g. their respective incomes, wealths, utilities, resources, liberties, rights, quality of life, and so on. The plurality of variables on which we can possibly focus (the *focal variables*) to evaluate interpersonal inequality makes it necessary to face, at a very elementary level, a hard decision regarding the perspective to be adopted. (Sen 1992, p. 20)

Pareto saw the distribution of income as a reflection of the natural distribution of abilities among persons, while Kuznets regarded its evolution as one of the characteristics of the process of economic growth; but they both agreed that the focal variable should be income. However, other dimensions of economic inequality are relevant in international comparisons. Earnings dispersion and differences in employment rates capture inequality in the labour market. Wealth may be seen as an indicator of the capacity to face adverse events or of the power to control the resources of the society. The standard of living is much influenced by non-monetary aspects, such as a person's health status or human capital – as stressed by the 'capability approach' advocated by Sen (1992).

In this article, the focal variable is taken to be income, the most common indicator of (current) economic resources in rich countries. Expenditure is an alternative variable often used, especially in less developed countries. The World Bank (2005, Table A2, pp. 280–1) reports income-based Gini indices for 22 of the 27 high-income economies for which the statistics are available vis-à-vis 20 of the 60 middle-income economies and only one of the 39 low-income economies. Mixing income-based and consumption-based statistics confounds international comparisons, as income tends to be more unequally distributed than expenditure – and to an extent that varies considerably from country to country (for example, World Bank 2005, Box 2.5, p. 38).

Wealth (net worth) is much more concentrated than income. Moreover, international comparisons of net worth are very problematic (Wolff 1996; Davies and Shorrocks 2000) as the assembling of cross-nationally comparable databases on household net worth is still in its infancy (Sierminska et al. 2006).

Inequality (International Evidence), Table 1 World Bank's estimates of inequality levels: Income and expenditure. Gini indices

Country	Year	Gini index	Income group
High-income economies			
Expenditure			
Taiwan	2000	0.24	HIC
Italy	2000	0.31	HIC
Israel	2001	0.35	HIC
Greece	1998	0.36	HIC
Income			
Finland	2000	0.25	HIC
Japan	1993	0.25	HIC
Sweden	2000	0.25	HIC
Belgium	2000	0.26	HIC
Denmark	1997	0.27	HIC
Norway	2000	0.27	HIC
Austria	1997	0.28	HIC
Germany	2000	0.28	HIC
Luxembourg	2000	0.29	HIC
Netherlands	1999	0.29	HIC
France	1994	0.31	HIC
Ireland	2000	0.31	HIC
Switzerland	1992	0.31	HIC
Australia	1994	0.32	HIC
Republic of Korea	1998	0.32	HIC
Canada	2000	0.33	HIC
United Kingdom	1999	0.34	HIC
Spain	2000	0.35	HIC
New Zealand	1997	0.37	HIC
United States	2000	0.38	HIC
Portugal	1997	0.39	HIC
Singapore	1998	0.43	HIC
Middle East and North Africa			
Expenditure			
Yemen	1998	0.33	LIC
Egypt	2000	0.34	LMC
Algeria	1995	0.35	LMC
Morocco	1998	0.38	LMC
Jordan	2002	0.39	LMC
Tunisia	2000	0.4	LMC
Iran	1998	0.43	LMC
South Asia			
Expenditure			
Pakistan	2001	0.27	LIC
Bangladesh	2000	0.31	LIC
India	1999/2000	0.33	LIC
Nepal	1996	0.36	LIC
Sri Lanka	2002	0.38	LMC
East Asia and Pacific			
Expenditure			
Mongolia	1998	0.3	LIC
Indonesia	2000	0.34	LMC
Lao PDR	1997/1998	0.35	LIC
Vietnam	2002	0.35	LIC
Cambodia	1997	0.4	LIC
Thailand	2002	0.4	LMC
China	2001	0.45	LMC
Philippines	2000	0.46	LMC
Income			
Malaysia	1997	0.49	UMC
Europe and Central Asia			
Expenditure			
Hungary	2002	0.24	UMC
Bosnia & Herzegovina	2001	0.25	LMC
Armenia	2003	0.26	LMC
Uzbekistan	2000	0.27	LIC
Bulgaria	2003	0.28	LMC
Romania	2002	0.28	LMC
Serbia & Montenegro	2003	0.28	LMC
Slovenia	1998	0.28	HIC
Croatia	2001	0.29	UMC
Kyrgyzstan	2002	0.29	LIC
Lithuania	2000	0.29	UMC
Belarus	2000	0.3	LMC
Kazakhstan	2003	0.3	LMC
Albania	2002	0.31	LMC
Poland	2002	0.31	UMC
Estonia	1998	0.32	UMC
Russian Federation	2002	0.32	UMC
Tajikistan	2003	0.32	LIC
Latvia	1998	0.34	UMC
Azerbaijan	2001	0.36	LMC
Macedonia	2003	0.36	LMC
Moldova	2001	0.36	LIC
Turkey	2002	0.37	UMC
Georgia	2002	0.38	LMC
Turkmenistan	1998	0.41	LMC
Income			
Czech Republic	1996	0.25	UMC

(*continued*)

I

Inequality (International Evidence), Table 1 (continued)

Country	Year	Gini index	Income group
Slovak Republic	1996	0.26	UMC
Ukraine	1999	0.29	LMC
Latin America and the Caribbean			
Expenditure			
Trinidad & Tobago	1992	0.39	UMC
Nicaragua	2001	0.4	LIC
Jamaica	2001	0.42	LMC
St. Lucia	1995	0.44	UMC
Peru	2000	0.48	LMC
Panama	2000	0.55	UMC
Income			
Venezuela	2000	0.42	UMC
Uruguay (urban)	2000	0.43	UMC
Guyana	1998	0.45	LMC
Costa Rica	2000	0.46	UMC
Dominican Republic	1997	0.47	LMC
Mexico	2002	0.49	UMC
El Salvador	2002	0.5	LMC
Argentina (urban)	2001	0.51	UMC
Chile	2000	0.51	UMC
Honduras	1999	0.52	LMC
Colombia	1999	0.54	LMC
Ecuador	1998	0.54	LMC
Paraguay	2001	0.55	LMC
Bolivia	2002	0.58	LMC
Guatemala	2000	0.58	LMC
Brazil	2001	0.59	LMC
Haiti	2001	0.68	LIC

Notes: Economies are classified by the World Bank according to 2004 per capita gross national income in the following income groups: low-income economies (LIC), $825 or less; lower-middle- income economies (LMC), $826–$3255; upper-middle income economies (UMC), $3256–$10,065; and high-income economies (HIC), $10,066 or more. (*Source*: World Bank (2005, Table A2, pp. 280–1))

Methodology

International comparisons of income inequality crucially depend on the underlying measurement assumptions. This has been known at least since Kravis (1962) and Kuznets (1963) and has received growing attention from the mid-1970s (for example, Atkinson 1974; Sawyer 1976; Lydall 1979). However, it was not until the assembling of the cross-nationally comparable database of the Luxembourg Income Study (LIS) that the impact of these assumptions was fully understood (Smeeding 2004). Differences in methodology arise in the definition of income, the choice of the recipient unit, the quality of underlying sources, the treatment of individual data (O'Higgins et al. 1990; Atkinson et al. 1995; Gottschalk and Smeeding 1997, 2000; Atkinson and Brandolini 2001).

Income definitions differ in comprehensiveness, as certain income sources like capital gains, imputed rents on owner-occupied dwellings, or home production may or may not be included. There are also widespread differences in the treatment of taxes (and social security contributions), as income may be taken before taxes, before taxes but after allowing for tax deductions, or after taxes. The definition of income may be augmented to include the imputed value of public in-kind benefits for education, health care and housing or to deduct indirect taxes. Moreover, income may be measured over a variety of time periods: the reference is often the year, but in some cases it is some 'current' period (for example, the most recent pay period for earnings in household surveys for the United Kingdom) and then the annual amount must be estimated.

The *reference unit* may be the household, the related or extended family, the tax unit, or the individual income earner. Information obtained from income tax records typically relates to the tax unit only, while sample surveys generally provide data for all members of a household. The total income may be adjusted for the size and the composition of the reference unit by dividing by an *equivalence scale.* Indeed, not adjusting income implies that the welfare achievable in a household with a certain income is independent of the number of its occupants. At the other extreme, taking income per capita amounts to an assumption that no economies of scale arise from cohabitation and that people do not differ in their needs. The *welfare unit* may be the person (person-weighted) or the household (household-weighted): in the former case the welfare indicator

represented by (equivalent) income is counted as many times as there are persons in the household, while in the latter it is counted only once. This welfare weighting is a separate issue from that of the equivalence scale: for instance, the European Commission (2002) typically reports statistics for the distribution of *equivalent disposable incomes* among *persons,* while the U.S. Census Bureau (2005) presents figures for the distribution of *unadjusted money incomes before taxes* among *households.*

Diversity in definitions is not the only factor that affects the comparability of income inequality statistics. There are also differences in the *nature of the data source*, the most important distinction being between sample surveys and administrative archives. Data may cover the whole population or only the household population, excluding people living permanently in institutions like boarding houses, nursing homes for the elderly, prisons, or military bases. Administrative data reflect the purposes for which they were collected. Even when sources have the same nature, they may considerably vary in quality, through differences in the response rate, the under-reporting of certain income components, or the coverage of the bottom and the top of the distribution. Lastly, significant differences can originate in the way data are processed. For example, the Gini index may be computed from micro-data or from observations grouped by income classes. When the ranking of observations is based on a variable different from that of concern, say before-tax income instead of after-tax income, measures of inequality are understated.

All these factors influence international comparisons of income inequality, as shown for instance by Buhmann et al. (1988) with regard to equivalence scales and by Smeeding et al. (1993) with regard to the inclusion of in-kind public benefits. These differences need to be kept in mind when making international comparisons. While perfect comparability is not achievable, it is important to raise the ratio of signal to noise by minimizing data and methodological differences across nations (Gottschalk and Smeeding 2000).

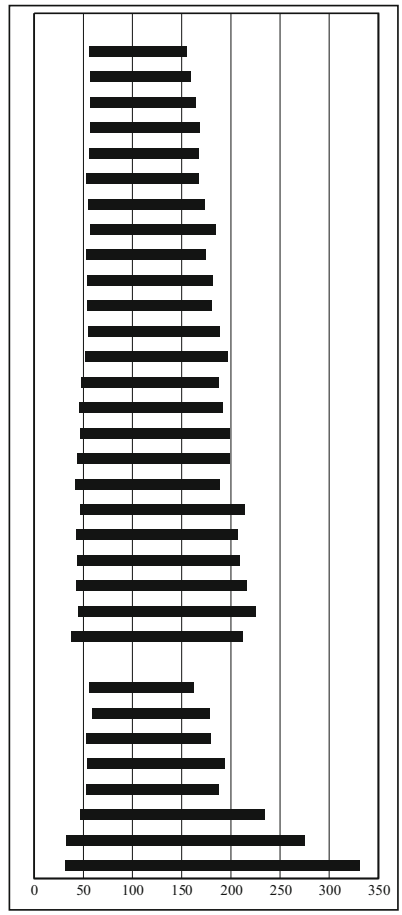
Relative Inequality Levels

Figure 1 compares the distribution of equivalent disposable income among persons in 32 nations for various years around the turn of the 21st century, or for the most recent year available in the LIS database. Disposable income is defined as the sum of wages, salaries and earnings from self-employment, cash receipts from property, private pension schemes, alimony and child support, public transfer payments (retirement pensions, family allowances, unemployment compensation, and welfare benefits) less income taxes and social security contributions. Observations are top- and bottom-coded in order to reduce the influence of anomalous income values. Total household income, the sum over all household members, is divided by a simple equivalence coefficient (the square root of the household size) and then attributed to each person in the household.

Figure 1 reports, for each country, the ratio to the median of the income of a person at the tenth percentile (P10 or 'low income') and a person at the 90th percentile (P90 or 'high income'). P10 and P90 provide some indication of how far below or above the middle of the distribution the poor and the rich are on the continuum of income. The ratio between P90 and P10, the 'decile ratio', is a measure of the gap between the rich and the poor. While these statistics refer to specific points of the distribution, the Gini index measures inequality across the entire distribution. For non-negative values, it varies between zero (perfect equality) and one (maximum inequality).

There is a wide range of income inequality among the nations of Fig. 1. The United States is an outlier among rich nations, and only Russia and Mexico, two middle-income economies, have higher levels of inequality. A low-income American at the 10th percentile in 2000 had an income that was only 37 per cent of the median income. By contrast, in most countries of central, northern and eastern Europe the income of the poor exceeded 50 per cent of the income of middle-income person; in the other English-speaking nations and in the southern European countries, plus Israel, it was above 40 per cent. Only in Russia and Mexico did the poor fare relatively

	P10 (Low income)	Length of bars represents the gap between high and low income individuals	P90 (High income)	P90/P10 (Decile ratio)	Gini index
High-income economies					
Denmark 2000	57		155	2.8	0.225
Norway 2000	57		159	2.8	0.251
Finland 2000	57		164	2.9	0.247
Sweden 2000	57		168	3.0	0.252
Netherlands 1999	56		167	3.0	0.248
Slovenia 1999	53		167	3.2	0.249
Austria 2000	55		173	3.2	0.260
Luxembourg 2000	57		184	3.2	0.260
Belgium 2000	53		174	3.3	0.277
Switzerland 2000	55		182	3.3	0.280
Germany 2000	54		180	3.4	0.275
France 2000	55		188	3.4	0.278
Taiwan 2000	52		196	3.8	0.296
Canada 2000	48		188	3.9	0.302
Japan 1992	46		192	4.2	0.315
Australia 2001	47		199	4.2	0.317
Italy 2000	45		199	4.5	0.333
Ireland 2000	41		189	4.6	0.323
United Kingdom 1999	47		215	4.6	0.343
Greece 2000	43		207	4.8	0.338
Spain 2000	44		209	4.8	0.340
Israel 2001	43		216	5.0	0.346
Portugal 2000	45		226	5.0	0.363
United States 2000	37		212	5.7	0.370
Middle-income economies					
Slovak Republic 1996	56		162	2.9	0.241
Czech Republic 1996	59		179	3.0	0.259
Romania 1997	53		180	3.4	0.277
Hungary 1999	54		194	3.6	0.295
Poland 1999	52		188	3.6	0.293
Estonia 2000	46		234	5.1	0.361
Russia 2000	33		276	8.4	0.434
Mexico 2000	32		331	10.4	0.491



Inequality (International Evidence), Fig. 1 The distribution of disposable income in 32 high- and middle-income economies. *Notes*: P10 and P90 are the ratios to the median of the tenth and 90th percentiles, respectively. Observations are bottom-coded at 1% of the mean of equivalent disposable income and top-coded at ten times the median of unadjusted disposable income. Incomes are adjusted for household size by the square-root equivalence scale. See note to Table 1 for the definition of high- and middle-income economy. (*Sources*: Authors' calculations from the Luxembourg Income Study database, as of 10 March 2007 (figures coincide with those reported in http://www.lisproject.org/keyfigures/ineqtable.htm) and the European Community Household Panel database, Waves 1–8, December 2003 for Portugal; statistics for Japan were computed according to the same methodology as all other figures by Tsuneo Ishikawa for Gottschalk and Smeeding (2000))

worse than in the United States. In Greece, Portugal, Spain, Israel as well as the United States and the United Kingdom the rich persons earn more than twice the national median incomes. In poorer countries the 90th percentile can also be very high in relative terms, for example in Mexico, Russia, and Estonia.

The countries in Fig. 1 fall into distinctive clusters. Inequality, as measured by the decile ratio, is least in Nordic countries, the Netherlands and the Czech and Slovak Republics with values of 3 or less. The two other Benelux countries (Belgium and Luxembourg), Central Europe (France, Switzerland, Germany, Austria, Slovenia) and three other Eastern European countries (Hungary, Poland, Romania) come next at 3.2–3.6. These precede four English-speaking nations (Canada, Australia, Ireland and the United Kingdom), which have decile ratios comprised between 3.9 and 4.6, and the southern European

countries (Italy, Spain, Greece and Portugal) and Israel, whose ratios fall between 4.5 and 5.0. Only the United States, Estonia, Mexico and Russia have values in excess of 5. With decile ratios around 4, the two Asian countries, Taiwan and Japan, are in an intermediate position.

Inequality differs much more across middle-income than high-income economies. While Estonia, Russia and Mexico show a very unequal distribution of income, the other five countries, all from eastern Europe, exhibit moderate or low levels of inequality. The shape of the income distribution was noticeably different even in the mid-1980s across these formerly planned economies, with Czechoslovakia showing the least inequality and the Soviet Union the highest (Atkinson and Micklewright 1992).

In Fig. 1 countries are arranged, within the two categories of high-income and middle-income, by the decile ratio, from lowest to highest. This country rank order does not need to coincide with that based on the other statistics reported: P10, P90 and the Gini index. For instance, Sweden shows the second highest P10 but the seventh lowest Gini index. This follows from the fact that the Swedish at the 90th percentile is less closer to the middle than the equivalent person in Denmark, Finland or the Slovak Republic. (-These differences should not be overstressed as they are small and likely to be within the bounds of sampling error.) The rankings of countries in international comparisons depends on which part of the distribution is analysed, for example, the bottom with P10 or the top with P90, or in the way single observations are weighted by a summary measure of inequality like the Gini index, or the Theil and Atkinson indices. Different summary measures may produce different results reflecting differences at the top and bottom of the distribution. More robust, but partial, rankings are obtained by comparing the entire distributions by 'Lorenz dominance', whereby inequality is assessed to be unequivocally higher in country *A* than in country *B* if the Lorenz curve of country *A* lies everywhere below that of country *B*, but no unambiguous conclusion is achieved if the two curves intersect. Although countries may switch their relative positions, indices are still in general highly correlated: for instance, the correlation between decile ratio and Gini index in Fig. 1 is 0.97. The basic patterns of international inequality are clear regardless of the measure of inequality employed.

Redistribution

Every nation's tax and benefit system reduces market income inequality, but not all are equally effective in doing so. The efficiency with which nations accomplish this redistribution may vary over time as well as space. A common measure of the level of redistribution is represented by the difference between the Gini index for market incomes, that is, before public transfers are added and taxes and social security contributions are deducted, and the Gini index for disposable incomes. This difference provides only a first estimate of the actual impact of public redistribution, as it ignores how market income inequality would be different if there were no taxes and benefits. Table 2 shows the extent of redistribution in 16 countries using LIS data.

In all nations disposable incomes are more equally distributed than market incomes, suggesting that the tax and benefit system narrows the overall distribution. On average, inequality falls by about a third, from a Gini index of 44 to one of 29 per cent. Cross-country variation in original inequality is wider than after redistribution: the Gini index ranges from 33 to 52 per cent for market incomes, and from 23 to 37 per cent for disposable incomes. The United States has the highest inequality of disposable incomes, although the dispersion of market incomes is on the high side but not far from most other countries; it is as high as in Germany and Australia and below the values recorded for the United Kingdom, Poland and Israel. The fact is that the percentage reduction in before-tax-and-benefit inequality in the United States is a mere 23 per cent. If we exclude Taiwan, where redistribution has a tiny impact, only Switzerland shows a reduction as low as the United States, but the Swiss start from a much more equal distribution and end with a Gini index below the average.

Inequality (International Evidence), Table 2 Gini indices of market income and disposable income in 16 countries (per cent)

Country	Year Gini index for market income [1]	Gini index for disposable income [2]	Absolute reduction [3] = [1] − [2]	Percentage reduction [4] = [3]/[1]
High-income economies				
Denmark	2000 42	23	20	47
Finland	2000 38	25	14	36
Netherlands	1999 39	25	14	36
Norway	2000 41	25	16	39
Sweden	2000 46	25	21	45
Germany	2000 48	28	21	43
Switzerland	2000 36	28	8	22
Taiwan	2000 33	30	3	9
Canada	2000 42	30	12	28
Australia	2001 48	32	17	34
United Kingdom	1999 51	34	17	33
Israel	2001 52	35	17	33
United States	2000 48	37	11	23
Middle-income economies				
Czech Republic	1996 44	26	18	41
Romania	1997 38	28	10	27
Poland	1999 50	29	21	41

Notes: Observations for disposable income are bottom-coded at 1% of the mean of equivalent disposable income and top-coded at ten times the median of unadjusted disposable income. Changes in disposable incomes due to bottom- and top-coding are entirely attributed to market incomes. Both market and disposable incomes are adjusted for household size by the square-root equivalence scale. (*Source*: Authors' calculations from the Luxembourg Income Study database, as of 10 March 2007)

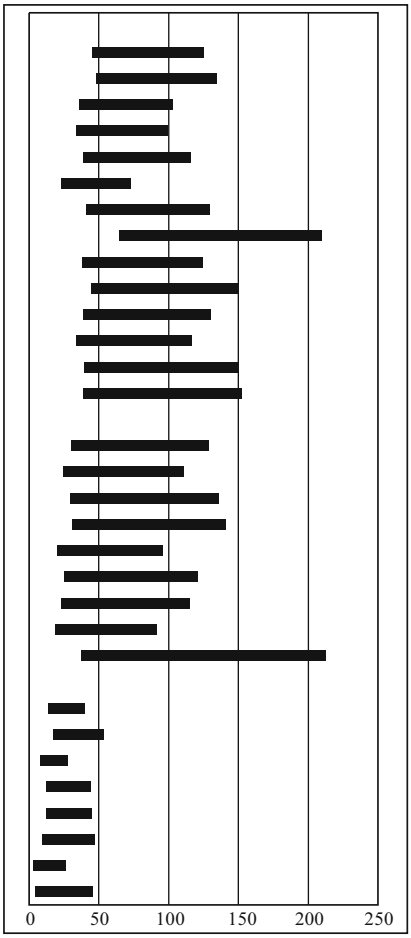
These percentage reductions are very consistent with the patterns of aggregate public spending. High-spending northern and central European nations have the highest degree of inequality reduction, from 36 to 47 per cent; the English-speaking Anglo-Saxon (excluding the United States) nations and Israel are next with 28 to 33 per cent reductions; the United States and Switzerland are, as just seen, at the bottom of the scale. The degree of redistribution in southern Europe is lower than in Ireland and the United Kingdom, especially if public pensions are not included among transfers, according to the EUROMOD estimates based on micro-simulations rather than the records of the original micro-data sources (Immervoll et al. 2005). The nations that redistribute the most are not necessarily those with the greatest degree of market income inequality: before-tax-and-benefit incomes in Finland and the Netherlands are far more equally distributed than in the United States.

Absolute Inequality Levels

The comparisons in Fig. 1 relate to *relative* inequality. The income of the poor at the tenth percentile is compared with the income of the person at the middle of the distribution in the same country. When average standards of living differ across nations, results may look quite different if comparisons are made in terms of *real* income, that is, the amount of goods that a certain income can purchase.

The statistics in Fig. 2 on real incomes in 2000 international dollars are derived by adjusting the original incomes by the national consumer price indices (CPI) and converting them by means of the purchasing power parities (PPP) for gross domestic product (GDP). The real P10 and P90 are then recomputed as a fraction of the US median real income. These comparisons are very rough indicators of differences in 'real living standards'. First, the conversion to real income across countries and

	Real P10 (Low income)	Length of bars represents the gap between high and low income individuals	Real P90 (High income)	P90/P10 (Decile ratio)	Real median
High-income economies					
Denmark 2000	45		125	2.8	80
Norway 2000	48		134	2.8	84
Finland 2000	36		103	2.9	63
Sweden 2000	34		99	3.0	59
Netherlands 1999	39		116	3.0	69
Slovenia 1999	23		73	3.2	44
Austria 2000	41		129	3.2	75
Luxembourg 2000	65		209	3.2	114
Belgium 2000	38		125	3.3	72
Switzerland 2000	45		150	3.3	82
Germany 2000	39		130	3.4	72
France 2000	34		117	3.4	62
Taiwan 2000	39		150	3.8	76
Canada 2000	39		152	3.9	81
Japan 1992				4.2	
Australia 2001	30		129	4.2	65
Italy 2000	25		111	4.5	56
Ireland 2000	30		136	4.6	72
United Kingdom 1999	31		141	4.6	66
Greece 2000	20		96	4.8	46
Spain 2000	25		121	4.8	58
Israel 2001	23		115	5.0	53
Portugal 2000	18		91	5.0	40
United States 2000	37		212	5.7	100
Middle-income economies					
Slovak Republic 1996	14		40	2.9	25
Czech Republic 1996	18		54	3.0	30
Romania 1997	8		27	3.4	15
Hungary 1999	12		44	3.6	23
Poland 1999	12		45	3.6	24
Estonia 2000	9		47	5.1	20
Russia 2000	3		26	8.4	10
Mexico 2000	4		46	10.4	14



Inequality (International Evidence), Fig. 2 The distribution of real disposable income in 32 high- and middle-income economies. *Notes*: Real P10 and P90 are the percentage ratios to the US median of the tenth and 90th percentiles, respectively; real median is expressed as a percentage ratio of the US median. Observations are bottom-coded at 1% of the mean of equivalent disposable income and top-coded at ten times the median of unadjusted disposable income. Incomes are adjusted for household size by the square-root equivalence scale. (*Sources*: Authors' calculations from the Luxembourg Income Study database, as of 10 March 2007, and the European Community Household Panel database, Waves 1–8, December 2003 for Portugal; statistics for Japan were computed according to the same methodology as all other figures by Tsuneo Ishikawa for Gottschalk and Smeeding (2000). Consumer price indices and purchasing power parity conversion factors from local currency units to international dollars are from International Monetary Fund (2006))

time is sensitive to the PPP and consumer price indices used. Second, the PPPs are computed for national accounts which are intrinsically different from survey data (Deaton 2005). For instance the ratios of total survey incomes to GDP aggregates vary considerably across these countries. Thus countries with surveys that capture less of national income appear to have much lower mean living standards than countries whose surveys or administrative records capture a larger share of that income. Third, it is questionable that the same conversion factor should be applied across the entire distribution. Lastly, real income does not account for goods and services such as education

and health care that are provided at different prices and under different financing schemes in different nations. As low-income citizens in some countries need to spend more out of pocket for these goods than do low- income citizens in other countries, their living standard is relatively lower than that measured by PPP-adjusted income.

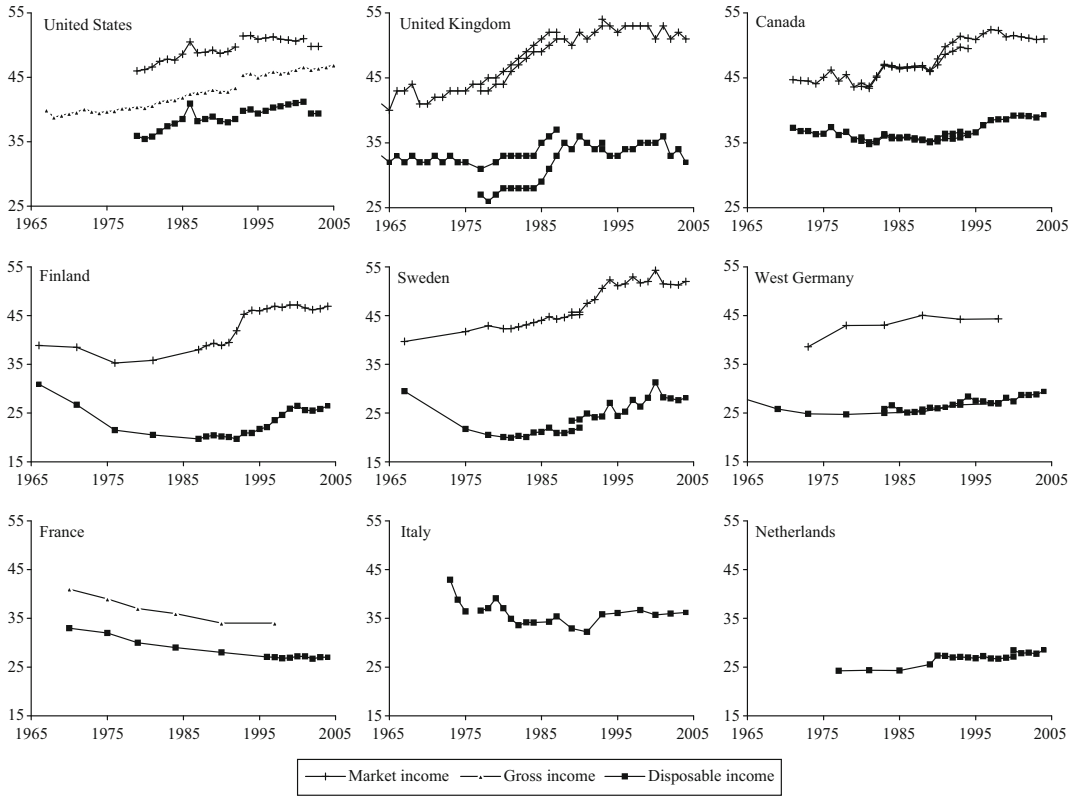
The living standard of the median German or Belgian appears to be 72 per cent of that of the median American; but the living standard of poor Germans and Belgians is just above that of their American counterparts, 38–39 per cent against 37 per cent of the US median. Low-income people in Denmark, Norway, Switzerland and, especially, Luxembourg are much better off than elsewhere. In all southern European countries but also, to a lesser extent, in Australia, Ireland and the United Kingdom, the living standards of low-income households were lower than in the United States. Of course, they are a great deal lower in all middle-income economies. At the other extreme, the rich Americans far surpass the rich in any other nation observed, save for the Luxembourgers.

Long-Run Trends in High-Income Economies

Movements of inequality over time follow irregular trajectories rather than smooth profiles, with more substantial changes often concentrated in few episodes (Atkinson 1997). Some causes are common to many countries, such as the spreading of skilled- biased technologies, the greater world economic integration, or the aging of population in more recent decades; some others are more specific to national experiences, typically changes in tax-and-benefit systems but also modifications in institutions such as wage setting policies. The evolution of inequality reflects the joint working of these factors, which sometimes balance out and sometimes reinforce each other, making it an arduous task to disentangle common trends from idiosyncratic variations. Moreover, changes in data collection and statistical methodology interrupt the continuity of time series. And so the interpretation of long-run movements needs to allow for the patchwork nature of the evidence.

The temporal patterns show some similarity in the United States and the United Kingdom, where inequality was considerably less in the 1940s than before the Second World War. It then moderately declined until the mid-1970s, when this trend abruptly reversed. But we have no consistent overall time series running this far back for other nations (see Gottschalk and Smeeding 2000, Figs. 6a and 6b, for the longer-term US and UK trends). The best we can do on a reasonably comparative basis is shown in Fig. 3, covering four decades from 1965 to 2005. (Estimates reflect national practices and are not to be compared across countries.) Indeed, the 1980s saw a substantial rise of inequality, more pronounced in Britain than in the United States, though the starting level was lower. In the 1990s the two nations parted: income distribution kept widening in the United States, while it broadly stabilized in the United Kingdom. Both Finland and Sweden experienced a fall in inequality until the early 1980s and then a modest rise afterwards, which has strengthened around the turn of the century. A tendency towards higher inequality followed by a period of stability seems to characterize the 1980s and the 1990s in the Netherlands and Norway as well as Australia and New Zealand. Canadian income inequality exhibited some variation but no clear trend from 1965 to the mid- 1990s, when it started to slowly rise. In the Federal Republic of Germany a sharp fall between 1962 and 1973 was followed by a period of stability and a modest rise over the 1990s. Income distribution narrowed in Italy from the 1970s to the 1980s; after a sharp widening at the beginning of the 1990s, there was virtually no change until 2004. In France alone, inequality steadily decreased between 1970 and the mid- 1990s, and remained stable afterwards.

In summary, national experiences vary and there is no one overarching common story. However, there was a general tendency for the disposable income distribution to narrow until the mid-1970s. Some increase in inequality was experienced by most nations in the 1980s to the 1990s, but its timing and magnitude differed widely across countries. In particular there was and is no regression to the mean pattern of change in the United States, which began with the most inequality in the late 1970s and has increasingly



Inequality (International Evidence), Fig. 3 Inequality trends in selected high-income economies (Gini index, per cent), 1965–2005. (*Source*: Authors' elaboration on national sources)

pulled away from the other nations through the early years of the twenty first century.

These observations mainly relate to disposable incomes. In the six countries for which data are available (Canada, the Federal Republic of Germany, Finland, Sweden, the United Kingdom and the United States), movements in market income inequality appear to be more synchronous, with a rise in the 1980s followed by stability thereafter. Changing public redistribution appears to be an important determinant of the time pattern of the inequality of disposable incomes. If we take, as before, the absolute difference between Gini indices, the redistributive impact of taxes and transfers initially increased and then stabilized or dropped in all countries except for the United States, where it remained quite stable over time. The United Kingdom stands out for having the most dramatic switch of regime, as in the early 1980s it apparently shifted from a situation not too different from the two Nordic countries to a model closer to that of the two North American countries. It is not possible to infer from this simple measure whether changes in redistribution are the automatic response of a progressive tax-and-benefit system to changes in the distribution of market incomes, or are instead the product of explicit policy choices (Atkinson 2004). Nevertheless, they confirm that a widening of the market income distribution need not result in a drastic increase in the inequality of disposable incomes.

See Also

- ▶ Gini Ratio
- ▶ Household Surveys
- ▶ Income Mobility
- ▶ Inequality (measurement)
- ▶ Lorenz Curve

▶ Pareto Distribution
▶ Pareto, Vilfredo (1848–1923)
▶ Redistribution of Income and Wealth
▶ Survey Data, Analysis of

Bibliography

Atkinson, A.B.. 1974. *The economics of inequality.* Oxford: Clarendon Press.

Atkinson, A.B.. 1997. Bringing income distribution in from the cold. *Economic Journal* 107: 297–321.

Atkinson, A.B.. 2004. Increased income inequality in OECD countries and the redistributive impact of the government budget. In *Inequality, growth, and poverty in an era of liberalization and globalization*, ed. G.-A. Cornia. Oxford: Oxford University Press.

Atkinson, A.B.., and A. Brandolini. 2001. Promises and pitfalls in the use of secondary data-sets: Income inequality in OECD countries as a case study. *Journal of Economic Literature* 39: 771–800.

Atkinson, A.B.., and J. Micklewright. 1992. *Economic transformation in Eastern Europe and the distribution of income*. Cambridge: Cambridge University Press.

Atkinson, A.B.., L. Rainwater, and T.M. Smeeding. 1995. *Income distribution in OECD countries: The evidence from the Luxembourg income study (LIS)*. Paris: OECD.

Atkinson, T., B. Cantillon, E. Marlier, and B. Nolan. 2002. *Social indicators: The EU and social inclusion*. Oxford: Oxford University Press.

Bresciani-Turroni, C. 1939. Annual survey of statistical data: Pareto's law and the index of inequality of incomes. *Econometrica* 7: 107–133.

Buhmann, B., L. Rainwater, G. Schmaus, and T.M. Smeeding. 1988. Equivalence scales, well-being, inequality, and poverty: Sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database. *Review of Income and Wealth* 34: 115–142.

Davies, J.B., and A.F. Shorrocks. 2000. The distribution of wealth. In *Handbook of income distribution*, ed. A.B. Atkinson and F. Bourguignon. Amsterdam: North-Holland.

Deaton, A. 2005. Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and Statistics* 87: 1–19.

European Commission. 2002. *European social statistics – Income, poverty and social exclusion: 2nd report.* Data 1994–1997. Luxembourg: Office for Official Publications of the European Communities.

Gottschalk, P., and T.M. Smeeding. 1997. Cross-national comparisons of earnings and income inequality. *Journal of Economic Literature* 35: 633–687.

Gottschalk, P., and T.M. Smeeding. 2000. Empirical evidence on income inequality in industrialized countries. In *Handbook of income distribution*, ed. A.B. Atkinson and F. Bourguignon. Amsterdam: North-Holland.

Immervoll, H., H. Levy, C. Lietz, D. Mantovani, C. O'Donoghue, H. Sutherland and G. Verbist. 2005. Household incomes and redistribution in the European Union: Quantifying the equalising properties of taxes and benefits. Working Paper No. EM9/ 05, EUROMOD.

International Monetary Fund. 2006. *World Economic Outlook, September 2006.* Washington, DC: International Monetary Fund. Online. http://www.imf.org/external/pubs/ft/weo/2006/02/pdf/weo0906.pdf. Accessed 1 Apr 2007.

Kravis, I.B. 1962. *The structure of income. Some quantitative essays*. Philadelphia: University of Pennsylvania Press.

Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45: 1–28.

Kuznets, S. 1963. Quantitative aspects of the economic growth of nations: 8. Distribution of income by size. *Economic Development and Cultural Change* 11: 1–80.

Lydall, H.F. 1979. Some problems in making international comparisons of inequality. In *Income inequality: Trends and international comparisons*, ed. J.R. Moroney. Lexington: Lexington Books.

O'Higgins, M., L. Rainwater, and T.M. Smeeding. 1990. *Poverty, inequality, and income distribution in comparative perspective: The Luxembourg Income Study (LIS)*. Washington, DC: Urban Institute Press, Wheatsheaf Books.

Pareto, V. 1897. *Cours d'économie politique.* Lausanne and Paris: Rouge & Pichon. Repr. in *Oeuvres Complètes,* ed. G.-H. Bousquet and G. Busino. Genève: Droz, 1964.

Sawyer, M. 1976. *Income distribution in OECD countries. OECD economic outlook. Occasional studies.*. Paris: OECD.

Sen, A.K. 1992. *Inequality reexamined*. Oxford: Clarendon Press.

Sierminska, E., A. Brandolini, and T.M. Smeeding. 2006. The Luxembourg wealth study – A cross-country comparable database for household wealth research. *Journal of Economic Inequality* 4: 375–383.

Smeeding, T.M. 2004. Twenty years of research on income inequality, poverty, and redistribution in the developed world: Introduction and overview. *Socio-Economic Review* 2: 149–163.

Smeeding, T.M., P. Saunders, J. Coder, S. Jenkins, J. Fritzell, A.J.M. Hagenaars, R. Hauser, and M. Wolfson. 1993. Poverty, inequality, and family living standards impacts across seven nations: The effect of noncash subsidies for health, education, and housing. *Review of Income and Wealth* 39: 229–256.

U.S. Census Bureau. 2005. *Income, poverty, and health insurance coverage in the United States: 2004.* Current population reports, P60–229.Washington, DC: U.S. Government Printing Office.

United Nations. 1951. National income and its distribution in under-developed countries. Statistical Papers, Series E, No. 3. New York: United Nations.

Wolff, E.N. 1996. International comparisons of wealth inequality. *Review of Income and Wealth* 42: 433–451.

World Bank. 2005. *World development report 2006. Equity and development*. New York/Oxford: Oxford University Press.

Inequality (Measurement)

F. A. Cowell

Abstract

This article provides an overview of the key issues in inequality measurement and shows how theoretical concepts are related to practical judgements. The principal axioms of distributional analysis are used to show the social-welfare underpinnings of standard ranking principles and to derive families of inequality indices. Recent developments that focus on income differences and reference income levels are examined.

Keywords

Atkinson inequality index; Coefficient of variation; Generalized Lorenz curve; Gini coefficient; Index numbers; Inequality (measurement); Lorenz curve; Pen's parade; Poverty; Poverty line; Principle of transfers; Ranking; Relative deprivation; Risk aversion; Social-welfare function; Theil, H.; Well-being

JEL Classifications

D63; E25; D3

Introduction

Inequality measurement is principally concerned with the comparison of personal income distributions in quantitative terms. In its modern form it is a branch of welfare economics although it clearly derives some of its intellectual heritage from statistics. It is distinct from the measurement of poverty and relative deprivation, although there are close analytical links to these topics. The motivation for taking the subject of inequality seriously is both analytical and practical: the principal concepts reviewed in this article are of concern to theoretical economists and are also used by policymakers. The subject touches on questions addressed by philosophers and by social scientists.

The type of issue under consideration can be illustrated by a simple example as depicted in Tables 1 and 2. These tables do not pretend to be the most general or the most suitable representation of the facts, but they are from an easily accessible source and give a convenient snapshot of what happened to the distribution of income in the United States over a span of about 30 years. From Table 1 it is clear that the bottom decile income experienced a 12.2 per cent growth over the period (in real terms) while the median grew by half as much again (18.3 per cent) and the top decile grew by almost four times as much (44.8 per cent). Table 2 describes what happened to the average incomes of particular *groups*. The average income of households in the bottom fifth of the distribution grew by just 10.1 per cent over the 30 years while the average income of households

Inequality (Measurement), Table 1 Quantile incomes and growth, United States 1974–2004

	q-quantile		
q	1974	2004	Growth
10%	$9,741	$10,927	12.2%
20%	$16,285	$18,500	13.6%
50%	$37,519	$44,389	18.3%
80%	$64,781	$88,029	35.9%
90%	$83,532	$120,924	44.8%
95%	$102,534	$157,185	53.3%

Note: Columns 2 and 3 give the upper limit of the bottom 10%, 20%,... of the population. Incomes are in 2004 dollars; the income-receiving unit is the household. *Source*: DeNavas-Walt et al. (2005, Appendix Table A3).

Inequality (Measurement), Table 2 Growth in average incomes for the five quintile groups and overall. United States, 1974–2004

	Average income		
Group	1974	2004	Growth
1st	$9,324	$10,264	10.1%
2nd	$23,176	$26,241	13.2%
3rd	$37,353	$44,455	19.0%
4th	$53,944	$70,085	29.9%
Top	$95,576	$151,593	58.6%
Overall	*$43,875*	*$60,528*	*38.0%*

Note: Columns 2, 3 give the average incomes of the bottom fifth, second fifth, ... Incomes are in 2004 dollars; the income-receiving unit is the household. *Source*: as for Table 1.

in the top fifth grew by 58.6 per cent. We return to the use of the concepts of quantiles and shares after introducing some of the technical equipment needed for analysing income distributions.

The thumbnail sketch suggests a substantial increase in inequality in the United States over the last quarter of the 20th century. But how much did inequality increase? In what ways can the impressionistic method of inequality comparisons suggested in the example be made precise and interpreted within the context of standard economic analysis? The purpose of this article is to provide a succinct overview of the role played by economic theory and other abstract principles in this class of problem and how to make sense of inequality comparisons such as those suggested in the example.

The sketch example in Tables 1 and 2 also illustrates some of the essential practicalities that have to be taken into account when implementing the principles of inequality measurement. Should we be focusing on households or individuals? What is the appropriate definition of income?

To follow the analysis there are few prerequisites: an understanding of utility and preference analysis is helpful but not essential to grasping the basic points that will be discussed.

Basics

Components of the Problem

The framework adopted here is not the most general approach, but one that is suitable for setting out the key ideas. We begin by considering the basic building blocks and then show how to assemble the constituent parts.

Income and Income Distribution

At the heart of the problem there is some scalar entity to be called 'income', but in practice this entity could be wealth, expenditure or some other economic quantity, the distribution of which is of particular interest. Income is distributed among a number of 'income receivers', which we will refer to as 'persons' (although the income receiver in practice may be a family or household). Suppose that there is a known number of income receivers n and that person i has income x_i. The *income distribution* is then simply the vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_n). \tag{1}$$

The set of all possible income distributions X is a subset of $\mathbb{R}^n$. The nature of X is going to depend in practice upon the precise definition of 'income': is it logically possible to have a zero value of x_i, for example? Or a negative value? As a working assumption we will take it that X consists of all vectors (1) such that $x_i \geq \underline{x}$ and leave open the specification of the lower bound $\underline{x}$ for particular instances of the inequalitymeasurement problem. Representations of the income distribution other than (1) will appear later in the discussion.

Indices

The topic of inequality measurement presumes that there is an inequality measure. An obvious interpretation of this is that there is some index I that, given a particular income distribution $\mathbf{x}$, yields a real number that is taken to be the amount of inequality exhibited by the distribution. In some ways the index I works like other well-known summary statistics of distributions, such as the mean

$$\mu(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

and the variance

$$\mathrm{var}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} [x_i - \mu(\mathbf{x})]^2. \tag{3}$$

Indeed, the variance itself is sometimes used as an inequality index, although it is more common to use a transformed version of it known as the *coefficient of variation*:

$$I_{\mathrm{CV}}(\mathbf{x}) := \frac{\sqrt{\mathrm{var}(\mathbf{x})}}{\mu(\mathbf{x})}. \tag{4}$$

One of the most commonly used indices in practice is the *Gini coefficient* defined as

$$I_{\mathrm{Gini}}(\mathbf{x}) := \frac{1}{2n^2\mu(\mathbf{x})}\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i - x_j|. \tag{5}$$

There are many more. However, rather than running through an exhaustive list of candidate

indices it is more useful to examine the principles that have usually been applied to construct indices; this we do by considering a priori what constitutes a 'suitable' inequality measure, the issue addressed in section "Axioms".

Ranking and Dominance

An apparently more flexible interpretation of the idea of inequality measurement is the idea of an inequality *ranking*. This is a partial ordering that picks up the general flavour of the kind of comparisons that we suggested in the introduction; the partial ordering is typically captured by a simple representational tool. Consider three of these.

The first of these tools is *Pen's parade* (named after the famous parable introduced by Pen 1974, ch. 3), which is simply the inverse of the empirical distribution function. To depict it let $x_{[i]}$ denote the ith smallest component in the vector (1) – the ith smallest income. Then take the collection of points

$$\left(\frac{1}{n}, x_{[i]}\right), \quad i = 1, 2, \ldots, n. \tag{6}$$

From this simple definition we can also introduce the idea of dominance. Take two distributions $\mathbf{x}'$ and $\mathbf{x}''$ in X where $\mathbf{x}' = (x'_1, x'_2, \ldots, x'_n)$ and $\mathbf{x}'' = (x''_1, x''_2, \ldots, x''_n)$. If it is true that $x'_{[i]} > x''_{[i]}$ for all $i = 1, 2, \ldots, n$ then we say that $\mathbf{x}'$ strictly Parade-dominates $\mathbf{x}''$.

The resulting graph plots income quantiles against population proportions: $x_{[i]}$ is the quantile corresponding to the bottom q per cent of the population where $q = 100\frac{i}{n}$. To illustrate the concept we use the information in Table 1 to produce a graph that looks like Fig. 1. In Pen's parable we imagine the whole population (seen as individuals rather than households) arranged in order on the [0,1] interval where each person's height has been altered in proportion to his/her income; the average-height income recipient in 1974 is located at position 0.57 in Fig. 1 (in other words, at a point 57 per cent along the horizontal axis the height of the Parade is exactly mean income) but in 2004 the average-height income recipient is located at position 0.61. Although the distribution of 2004 Parade-dominates the distribution in 1974, it is clear from Table 1 that overall the Parade shifted upwards in a lopsided fashion over the 30 years with the incomes of the very rich (95 per cent quantile) growing more than four times faster than those of the poor (10 per cent quantile); this shift suggests increased inequality over the period. However, by itself the Parade does not tell us much about inequality directly, although concepts closely related to it are widely used to characterize inequality comparisons. It is common to use *quantile ratios* for distributional comparisons: for example the popular '90–10 ratio' is given by $x_{[k]}/x_{[j]}$ where j and k are, respectively, the smallest integers satisfying $j/n \geq 10\%$ and $k/n \geq 90\%$: in the example above this ratio increased from 8.6 to 11.1. Furthermore, there is an important welfare-economic interpretation of the Parade that is discussed in section "Ranking Distributions" below.

For the second and third concepts we use the $x_{[i]}$ to derive the normalized income cumulations; for any $i = 1, 2, \ldots, a$ these are

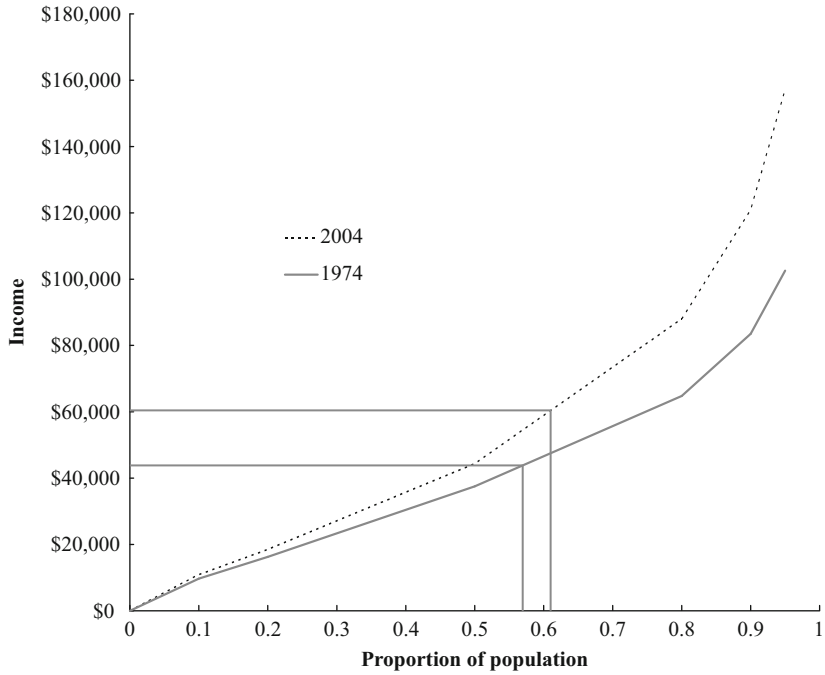
$$c_i := \frac{1}{n}\sum_{j=1}^{i} x_{[j]}. \tag{7}$$

Then the *generalized Lorenz curve* (GLC) is given by the graph of

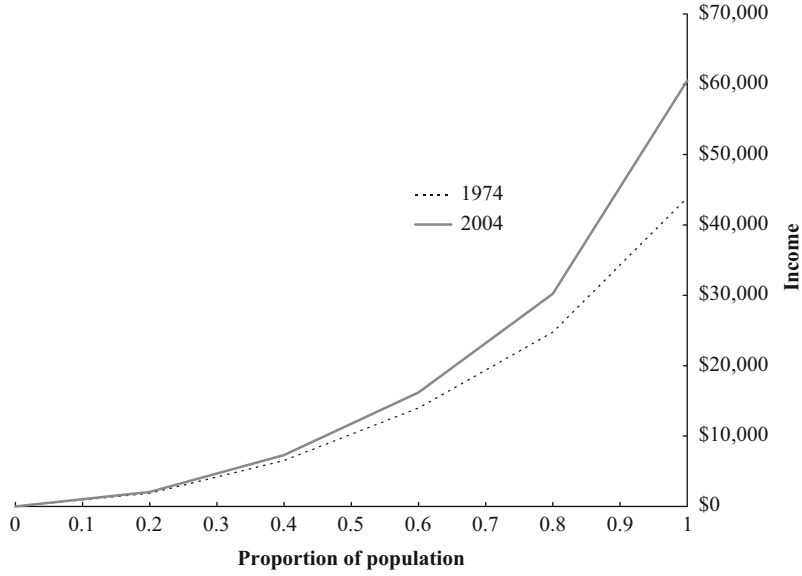
$$\left(\frac{i}{n}, c_i\right), i = 1, 2, \ldots, n. \tag{8}$$

Again we have a natural definition of dominance: for two distributions $\mathbf{x}'$ and $\mathbf{x}''$ in X if it is true that $c'_i > c''_i$ for all $i = 1, 2, \ldots, n$ then we say that $\mathbf{x}'$ strictly GLC-dominates $\mathbf{x}''$. For the example we used earlier the GLC is illustrated in Fig. 2, derived from Table 2. (Note that the definitions of Parade- and GLC-dominance can be extended to cases where the two distributions do not have the same number of incomes – this step makes use of the 'population principle' defined in section "Axioms". In some cases it is useful to consider the weak – non-strict – versions of the dominance criteria introduced here.)

The GLC plots the normalized income of the bottom $100q$ per cent of the population against q and, although the 2004 distribution GLC-dominates 1974, it is clear that over the period the growth of these group averages was not evenly distributed – the higher was q, the higher was the



Inequality (Measurement), Fig. 1 Parade diagram corresponding to Table 1 (*Source*: As for Table 1)



Inequality (Measurement), Fig. 2 Generalized Lorenz curve (*Source*: As for Table 1)

growth over 1974 to 2004. (This is easily inferred from Table 2: for example, the average income of the top 20 per cent grew almost six times as fast as the average income of the bottom 20 per cent.) Once again, although the GLC does not give information about inequality comparisons directly, there is an important welfare-economic interpretation (in section "Ranking Distributions"). In addition, a small modification of the GLC yields one of the central concepts of distributional analysis. Dividing c_i in (7)

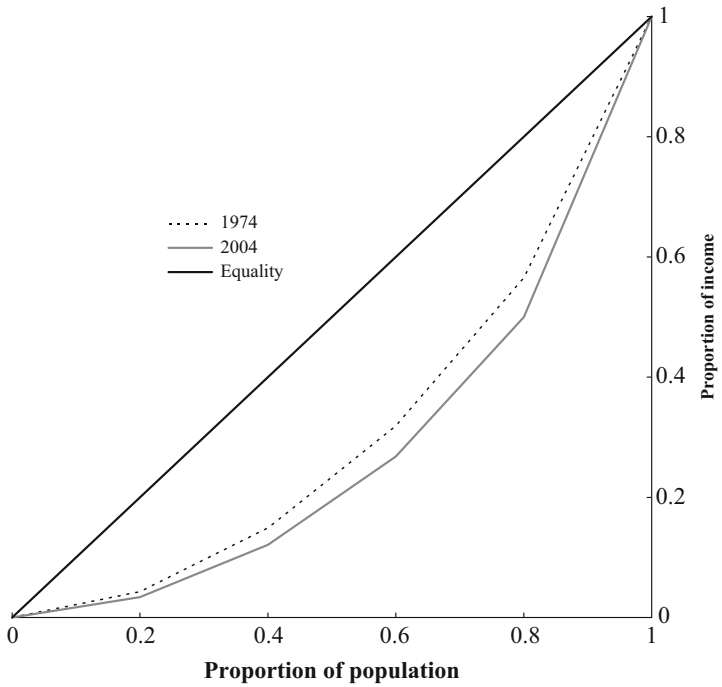
by the mean $\mu(\mathbf{x})$ gives the income share of the bottom $100\frac{i}{n}$ per cent of the population. The graph of the (population-proportion, income-share) pairs

$$\left(\frac{i}{n}, \frac{c_i}{\mu(\mathbf{x})}\right), i = 1, 2, \ldots, n \tag{9}$$

gives the *Lorenz curve*. Also, for two distributions $\mathbf{x}'$ and $\mathbf{x}''$, if it is true that, $c_i'/\mu(\mathbf{x}') > c_i''/\mu(\mathbf{x}'')$ for all $i = 1, 2, \ldots, n-1$ then we say that $\mathbf{x}'$ strictly Lorenz-dominates $\mathbf{x}''$. In the case of the example using US data this is illustrated in Fig. 3: the Lorenz curve plots the income share of the bottom $100q$ per cent of the population against q and the diagonal line depicts a hypothetical distribution of perfect equality. (Take the area trapped between the Lorenz curve and the equality diagonal. Using (7) and (9) we can show that the ratio of this area to the area of the whole triangle is given by the weighted sum $\sum_{i=1}^{n} \kappa_i x_{[i]}$ where the weights are $\kappa_i := [2i - 1 - n]/[n\mu(\mathbf{x})]$. This is exactly the Gini coefficient (5).) It is clear that for each q the share was smaller in 2004 than it was in 1974 – the 1974 distribution Lorenz-dominates that for 2004. This simple intuitive notion of greater inequality conforms exactly with a fundamental principle to be explained below.

Axioms

An inequality index I is in some ways like a utility function in consumer theory: it is a representation of an inequality ordering on the members of X and is usually taken to be continuous and ordinal – although there is often a 'natural' cardinal representation of a particular index, a formal argument for one representation rather than another is not usually provided (why not use the square or the log of the Gini coefficient?). Ordinality is sufficient for making comparing income distributions, the primary task of inequality analysis. Axioms are essentially formal statements of the principles of assessment that are used to give meaning to the ordering represented by I. The treatment here does not claim to generality; rather, it focuses on those principles that are central to modern approaches to inequality. Rather than presenting the axioms as formal statements, however, it is more useful here to introduce the underlying key principles discursively.



Inequality (Measurement), Fig. 3 Lorenz curve (*Source*: As for Table 1)

Assume that everywhere in the following discussion the vector **x** in (1) is any arbitrary member of the set X.

- First, it seems reasonable that the labelling of the components of **x** be irrelevant: it does not matter which income receiver gets which income. This means that I has the *symmetry* property:

$$\begin{aligned} I(x_1, x_2, \ldots, x_n) &= I(x_2, x_1, \ldots, x_n) \\ &= I(x_3, x_1, \ldots, x_n) = \ldots \end{aligned} \tag{10}$$

- We will always assume that this holds and we may therefore adopt the convention that incomes have been labelled such that $x_1 \leq x_2 \leq \ldots \leq x_{n\text{-}1} \leq x_n$.
- Second, we need some coherent way of characterizing inequality in differentsized populations. Perhaps the most obvious assumption is that simple replications of an income vector (1) leave inequality unchanged. This is the *population principle*:

$$\begin{aligned} I(x_1, x_2, \ldots, x_n) &= I(x_1, x_1, x_2, x_2, \ldots, x_n, x_n) \\ &= I(x_1, x_1, x_1, x_2, x_2, x_2, \ldots, x_n, x_n, x_n) = \ldots \end{aligned} \tag{11}$$

Taken in conjunction with symmetry this allows one to represent distributions purely in terms of a distribution function.

A key assumption that is commonly invoked focuses on the effect on inequality of a hypothetical small income transfer. Suppose $x_i < x_j$ and consider some positive number δ such that $x_i - \delta \geq \underline{x}$, then the *principle of transfers* (Dalton 1920) requires that:

$$\begin{aligned} &I(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n) \\ &\times < I(x_1, \ldots, x_i - \delta, \ldots, x_j + \delta, \ldots, x_n) \end{aligned} \tag{12}$$

  - a poorer-to-richer income transfer will always increase inequality.

As a counterpart to the assumption relating to different sizes of population (eq. 11) it is useful to have an assumption relating to different amounts of total income. The standard assumption is that of *scale independence*. This requires that, for any scalar $\lambda > 0$:

$$I(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) = I(x_1, x_2, \ldots, x_n) \tag{13}$$

  - double all incomes or halve all incomes and inequality is left unaltered.

An alternative assumption that is sometimes used is *translation independence*. Take any real number δ such that $x_1 + \delta \geq \underline{x}$; then

$$\begin{aligned} &I(x_1 + \delta, x_2 + \delta, \ldots, x_n + \delta) \\ &= I(x_1, x_2, \ldots, x_n) \end{aligned} \tag{14}$$

  - add or subtract one dollar from every income and inequality is left unaltered.

Clearly this brief list raises some important questions. Why use these particular axioms? Some of them appear to be quite strong; for example, although scale independence seems attractive if the 'incomes' x_i here are measured in dollars and we consider just dividing through by some rate of exchange so as to work with incomes in some other monetary units, it may seem less attractive if we want to consider the impact on inequality of redistribution policies at different stages of economic growth: a rearrangement of income shares that constitutes a reduction in inequality in a low-income society might not be considered as a reduction in inequality if the whole population is prosperous. Furthermore, the axioms captured by Eqs. (10), (11), (12), and (13), for example, are satisfied by both (4) and (5) as well as other important classes of inequality measures; on the other hand, the axioms captured by Eqs. (10), (11), (12) and (14) are satisfied by (3) and another rich class of inequality measures. Following on from this question, what more is required to get a specific index or well-defined family of indices that is both theoretically appropriate and practical to implement?

To answer this we need to be precise about what it means to say that one distribution is more unequal than another and the intellectual basis used for making such comparisons. The meaning of inequality can be further clarified through one of several routes: this article will analyse three of these in turn, namely, social welfare, decomposition, income differences.

Social Welfare and Inequality

The welfare-economic approach to the subject starts from the position that inequality is about 'illfare' – the opposite of welfare. If we adopt this approach then the definition of inequality follows almost immediately. The idea is similar to the conventional measurement of economic waste and the basis for a simple model can be laid with only a little more theorizing.

The *social-welfare function* (SWF) is a real-valued function W defined on the space of distributions X. The social welfare associated with a particular income distribution (1), given by

$$W(x_1, x_2, \ldots, x_n), \tag{15}$$

is to be interpreted as follows: suppose we are given a specific SWF $W(\,)$ and that for two separate income distributions $\mathbf{x}'$ and $\mathbf{x}''$ we have $W(\mathbf{x}') > W(\mathbf{x}'')$; then social welfare associated with the distribution $\mathbf{x}'$ is higher than the social welfare associated with the distribution $\mathbf{x}''$. In principle W is an ordinal function so that the scale of measurement of welfare levels can be subjected to arbitrary monotonic-increasing transformations.

This basic specification raises a number of important questions:

Why express social welfare as a function of income? Income defined how?

What particular form should W take?

What is the relation between the functions I and W?

The answer to the first question helps to pin down the relationship between inequality measurement as conventionally practised and standard welfare economics – see section "Introduction". The answers to the last two questions will determine the form of a class of inequality measures and permit us to establish some important welfare-economic results: these are addressed in sections "Basics" and "Social Welfare and Inequality".

Welfare and Income

We need to rectify a point that was fudged in the discussion of the US example: how to do the trick of passing from a distribution of dollar income among households to a standard welfare analysis that is typically concerned with the levels of economic well-being of individuals. The standard approach is as follows. We require a method of appropriately capturing the relationship between the living standard that is attainable by an individual and the income that he/she is presumed to have access to within the household. This is conventionally done by defining a function $v(.)$ that has as its argument a list of non-income attributes $\mathbf{a}$ that might include household size, age and sex of household members and health status; $v(\mathbf{a})$ determines the number of *equivalent adults* in the household with attributes $\mathbf{a}$ such that

$$x = \frac{y}{v(\mathbf{a})}, \tag{16}$$

where y is nominal income and x is *equivalized income* that is taken to be comparable across different household types. Note that the equivalization function v is typically specified as independent of income although this simplification is not essential; of course, the way in which the function v is determined – from ethical considerations or econometric studies – is an important issue in its own right, but one that lies outside the present discussion. The function v transforms a distribution of dollar incomes among n households

$$\mathbf{y} = (y_1, y_2, \ldots, y_n) \tag{17}$$

into a distribution of equivalized incomes by households given by (1). In order to complete the welfare interpretation we need to recognize that social-welfare considerations are usually represented in terms of individuals rather than households, and so, for example, households consisting of couples should receive more weight in social-welfare evaluations than households consisting of single individuals. Therefore, if the income-receiving units consist of households of differing size, we might want to represent this by introducing a corresponding set of population weights w_i for the observations, so that the distribution becomes an ordered list of pairs:

$$((w_1, x_1), (w_2, x_2), \ldots, (w_n, x_n)) \tag{18}$$

I

where w_i is the number of persons in household i divided by the number of persons in the whole population. There is little analytical complication in using (18) rather than (1) as a representation of the distribution of equivalized incomes by individuals. Typically it is just a matter of a minor redefinition of formulas for inequality measures and the like: for example, the coefficient of variation (4) would now be written

$$\sqrt{\sum_{i=1}^{n} w_i \left[\frac{x_i}{\mu} - 1\right]^2} \tag{19}$$

where μ is the appropriately redefined mean $\sum_{i=1}^{n} w_i x_i$. (More generally: all measures that can be written in the form $\Phi\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i), \mu\right)$ just need to be rewritten in the form $\Phi\left(\sum_{i=1}^{n} w_i\varphi(x_i), \mu\right)$. A similar modification applies to the Gini coefficient.)

However, having introduced this important theoretical qualification we will now neglect it – for expositional purposes it is convenient to assume (*a*) that the population consists of isolated individuals that are identical in every relevant respect other than income and (*b*) that income appropriately represents individual welfare. So, from here on, i indexes individuals or households and the distinction between x and y is dropped.

Social Welfare and Inequality Measures

The idea of the SWF was introduced without discussing specific properties of the function W. Some properties must be imposed on W if we require there to be a specific relationship between social welfare and inequality and we impose specific assumptions on the function I. However, in addition it is particularly important to be explicit about how W should respond to an increase in one or more incomes. This is the usual principle that is applied:

Suppose we consider any income distribution $(x_1, x_2, \ldots, x_n)$ and some positive number δ. Then *monotonicity* requires that:

$$W(x_1, x_2, \ldots, x_i + \delta, \ldots, x_n) > W(x_1, x_2, \ldots, x_i, \ldots, x_n) \tag{20}$$

On the assumption that monotonicity holds and that W is a continuous function, the SWF can itself be used to derive a family of inequality measures. There are several ways of doing this, but a standard approach is to represent social welfare using a money metric: we can always do this in view of the ordinal nature of W and the requirement that it be monotonic and continuous. The *equally distributed equivalent* (EDE) income is a real number ξ such that for any $(x_1, x_2, \ldots, x_n)$ in X:

$$W(\xi, \xi, \ldots, \xi) = W(x_1, x_2, \ldots, x_n). \tag{21}$$

(Note that monotonocity is unnecessarily strong for this step: for example one could define ξ in cases where one required only that W is increasing if *all* incomes are increased by δ, not just if *some* income is increased by δ. However, the assumption of monotonicity is useful for other results that follow.)

Clearly the relationship (21) can be used to derive EDE as a function of the income distribution, $\xi(\mathbf{x})$ and the function $\xi(\ )$ is a valid way of representing social welfare.

Suppose we require that the principle of transfers apply to W; this by analogy with (12) means that a mean-preserving poorer-to-richer income transfer will *decrease* social welfare. Then it is always true that $\xi(\mathbf{x}) \leq \mu(\mathbf{x})$ and the normalized gap between ξ and μ provides a natural basis for an inequality index

$$1 - \frac{\xi(\mathbf{x})}{\mu(\mathbf{x})}. \tag{22}$$

It is clear that this index is bounded between zero and 1 and that if there were perfect equality then we would have $\xi(\mathbf{x}) = \mu(\mathbf{x})$ and inequality in (22) would be zero.

Furthermore, if the scale-independence property (13) is also satisfied, then EDE income takes the form of a generalized mean:

$$\xi(\mathbf{x}) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}, \varepsilon > 0 \tag{23}$$

and (22) gives the class of *Atkinson indices*:

Inequality (Measurement), Table 3 Inequality indices for the example in Table 1

	1974	2004
$I_{\mathrm{A}}^{0.25}$	0.067	0.097
$I_{\mathrm{A}}^{0.5}$	0.134	0.190
$I_{\mathrm{A}}^{0.75}$	0.207	0.286
I_{A}^{1}	0.297	0.418
I_{Gini}	0.395	0.466
I_{GE}^{0}	0.352	0.542
I_{GE}^{1}	0.267	0.406

$$I_A^{\varepsilon}(\mathbf{x}) := 1 - \left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{\mu(\mathbf{x})}\right]^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}. \qquad (24)$$

(The limiting forms of (23) and (24) as $\varepsilon \to 1$ are, respectively,

$$\xi(\mathbf{x}) = \exp\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\log(x_i)\right) \text{ and } I_A^1(\mathbf{x}) = 1 - \exp\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\log(x_i)\right)/\mu(\mathbf{x}).)$$

The number ε– the degree of (relative) inequality aversion – is a parameter that characterizes individual members of the class of inequality measures. For any given unequal income distribution, the larger is ε the larger is the Atkinson inequality index – there is an example of this in Table 3 above. There is a close analogy with a class of risk indices in the case of constant relative risk aversion. This is unsurprising since this approach was explicitly founded on the formal similarity between distributional comparisons in terms of inequality and of risk (Atkinson 1970).

If, instead of the scale-independence property, we required I to satisfy translation independence (14), then we would obtain a different class of indices

$$I_{\mathrm{K}}^{\beta}(\mathbf{x}) := \frac{1}{\beta}\log\left(\frac{1}{n}\sum_{i=1}^{n}e^{\beta[x_i-\mu(\mathbf{x})]}\right) \qquad (25)$$

where $\beta > 0$ is a sensitivity parameter indexing members of the class (Kolm 1976). The connection of (25) with constant absolute risk aversion is evident. (One could also use 'regularity' assumptions other than scale- or translation-independence – see Bossert and Pfingsten 1990.)

Ranking Distributions

As noted earlier, there are important results available about welfare and inequality comparisons that do not require the usage of specific indices. They follow from standard first- and second-order dominance results that are familiar from finance and other disciplines. Take the special class of *additive* welfare functions where W in (15) can be written in the form $\sum_{i=1}^{n} u(x_i)$ for some function u. If W is additive and satisfies the monotonicity axiom, then u must be a strictly increasing function; if, furthermore, W satisfies the principle of transfers then u must be strictly concave. Then the following powerful results are available for any two distributions $\mathbf{x}'$ and $\mathbf{x}'' \in X$:

- *First-order*: $\mathbf{x}'$ strictly Parade-dominates $\mathbf{x}''$ if and only if $W(\mathbf{x}') > W(\mathbf{x}'')$ for any additive W that satisfies the principle of monotonicity.
- *Second-order*: $\mathbf{x}'$ strictly GLC-dominates $\mathbf{x}''$ if and only if $W(\mathbf{x}') > W(\mathbf{x}'')$ for any additive W that satisfies monotonicity and the principle of transfers (Shorrocks 1983).

(From the statement of these results it is clear that the Parade-dominance and GLC-dominance criteria are formally equivalent to first-order and second-order stochastic dominance in the analysis of probability distributions – see stochastic dominance.)

A version of the second-order result applies to the conventional Lorenz curve and it accords with the intuitive argument presented in the introduction. Take the class of SWFs that satisfy the principle of transfers (they do not have to be additive). Then, for two distributions $\mathbf{x}'$ and $\mathbf{x}''$ that have the same mean, the statement '$W(\mathbf{x}') > W(\mathbf{x}'')$ for any W in this class' is true if and only if $\mathbf{x}'$ strictly Lorenz-dominates $\mathbf{x}''$. Furthermore, under these circumstances for any inequality index I that satisfies the principle of transfers it must be the case that $I(\mathbf{x}') < I(\mathbf{x}'')$. The implication of this is that all inequality measures that satisfy the principle of transfers 'go the same way' if one distribution

Lorenz-dominates the other. This is illustrated in Table 3 (which again uses the distribution of household income by households). Rows 1 to 4 give the results for the Atkinson indices: notice that in each case measured inequality is closer to 1 (the maximum) the higher is the degree of inequality aversion. The indices in the last two rows of Table 3 are discussed in the next section.

Decomposition

The axioms discussed in section "Axioms" induced some structure on inequality measures. By introducing the idea of *decomposing* inequality we can impose more structure and thereby obtain a useful class of indices. There are two principal types of decomposition: by subgroups of the population (regions, age groups,...) and by components of income (labour income, income from capital,...). Here we focus just on the population-subgroup issue.

Imagine that the population of n persons can be partitioned into a collection of m groups so that any individual falls into just one of these m groups. Each group j could be considered as a sub-population of size n_j in its own right (where $\sum_{j=1}^{m} n_j = n$) and one could compute inequality within this subpopulation as

$$\iota_j = I(\mathbf{x}_j) \tag{26}$$

where $\mathbf{x}_j$ is the income distribution consisting of just the members of subgroup j. The essence of the decomposition problem is to represent inequality overall as a function of inequality in each group $j = 1, \ldots, m$

$$I(\mathbf{x}) = F(\iota_1, \iota_2, \ldots, \iota_m; \pi_1, \ldots, \pi_m, s_1, \ldots, s_m) \tag{27}$$

where F is an aggregation function and the terms after the ';' show that aggregation may depend on the groups' shares of the population $\pi_j := n_j/n$ and the groups' shares of total income $s_j := n_j\mu(\mathbf{x}_j)/n\mu(\mathbf{x})$. A consistency requirement on (27) is that, if the income distribution within subgroup j changes so as to increase ι_j in (26), all other things remaining the same, then inequality overall should increase. Insisting on this requirement on F for all logically possible partitions induces a type of separability on the function $I(\,)$ so that the index must be of the general form mentioned just after Eq. (19) above. If we also require that scale-independence hold, then the inequality index must take the specific form

$$I^{\alpha}_{\mathrm{GE}}(\mathbf{x}) = \frac{1}{\alpha^2 - \alpha}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{\mu(\mathbf{x})}\right]^{\alpha} - 1\right] \tag{28}$$

or some monotonic transform of it, where α is a real number. The 'GE' used in the labelling of (28) stands for the *generalized entropy* class, which is a generalization of the two indices introduced by Theil (1967). (Theil's two indices are those corresponding to the special forms in the cases $\alpha = 0, 1 : I^{0}_{\mathrm{GE}}(\mathbf{x}) := -\frac{1}{n}\sum_{i=1}^{n}\log(x_i/\mu(\mathbf{x}))$ and $I^{1}_{\mathrm{GE}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}[x_i/\mu(\mathbf{x})]\log(x_i/\mu(\mathbf{x}))$. The values of these indices for the US example are given in the last two rows of Table 3.) The α in (28) is a parameter that characterizes different members of the GE class: a high positive value of α yields an index that is very sensitive to income transfers at the top of the distribution; specifying a negative value will produce an index that is sensitive to income transfers among the poor. (There is a functional relationship between the class (24) and the class (28). For any $\alpha < 1$ we have $I^{\varepsilon}_{A}(\mathbf{x}) = 1 - \left[1 + \alpha[\alpha - 1]I^{\alpha}_{\mathrm{GE}}(\mathbf{x})\right]^{1/\alpha}$ where $\varepsilon = 1 - \alpha$.)

Income Differences

The third way forward from the basic argument outlined in section "Axioms" focuses on fundamental income differences. This is one of the key ways in which one can motivate usage of the very well-known inequality indices mentioned in section "Indices". The variance and the coefficient of variation (4) can be thought of as a representation of the averaged squared difference between each income x_i and the mean. A compelling argument for the Gini coefficient is that it is the (normalized) expected value of the absolute difference between any two randomly selected incomes in the population.

However, there are other types of income difference that are of special relevance to inequality measurement. Just as some poverty indices can be characterized as a kind of average distance of individual incomes from a reference income level – the poverty line (many poverty indices can be written in the form $\frac{1}{n}\sum_{i=1}^{n} p(z - x_i)$ where z is the poverty line and $p(\ )$ is a non-decreasing function that is zero for all $x_i \geq z$) – so also some inequality measures use the idea of a reference level income. In the case of inequality the reference income level has been suggested as either that of the best-off person in society, or of the average income of all those who are better off than any given person i (Temkin 1993). In each of these cases application of standard axioms about the structure of inequality orderings leads to a class of inequality indices that bears a functional similarity to poverty indices and to indices of relative deprivation (Cowell and Ebert 2004).

Implementation

The practical issues associated with the exposition of the example in Tables 1 and 2 highlight some of the problems in implementing inequality measures and associated tools – the definition of income, income receiver, and so on. Given the way in which income data are usually obtained, issues of sampling and measurement error usually need to be treated carefully. Furthermore, the special nature of income and wealth distributions and the sensitivity of inequality indices to very high or very low incomes usually require that particular attention be paid to the problem of outliers. Finally, it should be noted that it is still sometimes the case that the data required for estimating inequality indices are made available only in grouped form rather than as microdata so that special techniques may be required for interpolation within income intervals and for modelling the tails of the distribution.

Further Reading

For the welfare-economic issues, see Atkinson (1983) and Sen and Foster (1997). For literature surveys see Cowell (2000, 2007) and Lambert (2001).

See Also

- ▶ Poverty
- ▶ Stochastic Dominance

Acknowledgments My thanks go to Yoram Amiel, Tony Atkinson, Sanghamitra Bandyopadhyay, Kristof Bosmans, Udo Ebert, Giovanni Ko, Peter Lambert and Abigail McKnight, who made helpful comments on an earlier draft.

Bibliography

Atkinson, A.B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.
Atkinson, A.B. 1983. *The economics of inequality*. 2nd ed. Oxford: Clarendon Press.
Bossert, W., and A. Pfingsten. 1990. Intermediate inequality: Concepts, indices and welfare implications. *Mathematical Social Sciences* 19: 117–134.
Cowell, F.A. 2000. Measurement of inequality. In *Handbook of income distribution*, ed. A.B.. Atkinson and F. Bourguignon. Amsterdam: North Holland.
Cowell, F.A. 2007. *Measuring inequality*. 3rd ed. Hemel Hempstead: Oxford University Press.
Cowell, F.A., and U. Ebert. 2004. Complaints and inequality. *Social Choice and Welfare* 23: 71–89.
Dalton, H. 1920. Measurement of the inequality of incomes. *Economic Journal* 30: 348–361.
DeNavas-Walt, C., B.D. Proctor, and C.H. Lee. 2005. U. S. Census Bureau, Current Population Reports, P60–229, *Income, poverty, and health insurance coverage in the United States: 2004*. Washington, DC: U. S. Government Printing Office. Online. Available at http://www.census.gov/prod/2005pubs/p60-229.pdf. Accessed 24 Nov 2006.
Kolm, S.-C. 1976. Unequal inequalities I. *Journal of Economic Theory* 12: 416–442.
Lambert, P.J. 2001. *The distribution and redistribution of income*. 3rd ed. Manchester: Manchester University Press.
Pen, J. 1974. *Income distribution*. 2nd ed. London: Allen Lane, Penguin Press.
Sen, A.K., and J.E. Foster. 1997. *On economic inequality*. 2nd ed. Oxford: Clarendon Press.
Shorrocks, A.F. 1983. Ranking income distributions. *Economica* 50: 3–17.
Temkin, L.S. 1993. *Inequality*. Oxford: Oxford University Press.
Theil, H. 1967. *Economics and information theory*. Amsterdam: North Holland.

I

Inequality Between Nations

François Bourguignon

Abstract

Inequalities between nations are large but determining their magnitude and whether they have increased or decreased over time is not a simple matter. The conclusions that one reaches depend on the concept of inequality that one uses and on the point of view that one adopts.

Keywords

Child mortality; Educational attainment; Gini ratio; Global inequality; Globalization; Household surveys; Income mobility; Inequality between nations; Inter-country inequality; International inequality; Nutrition; Power; Purchasing power parity; Value judgements; Voice; Well-being

JEL Classifications

O4

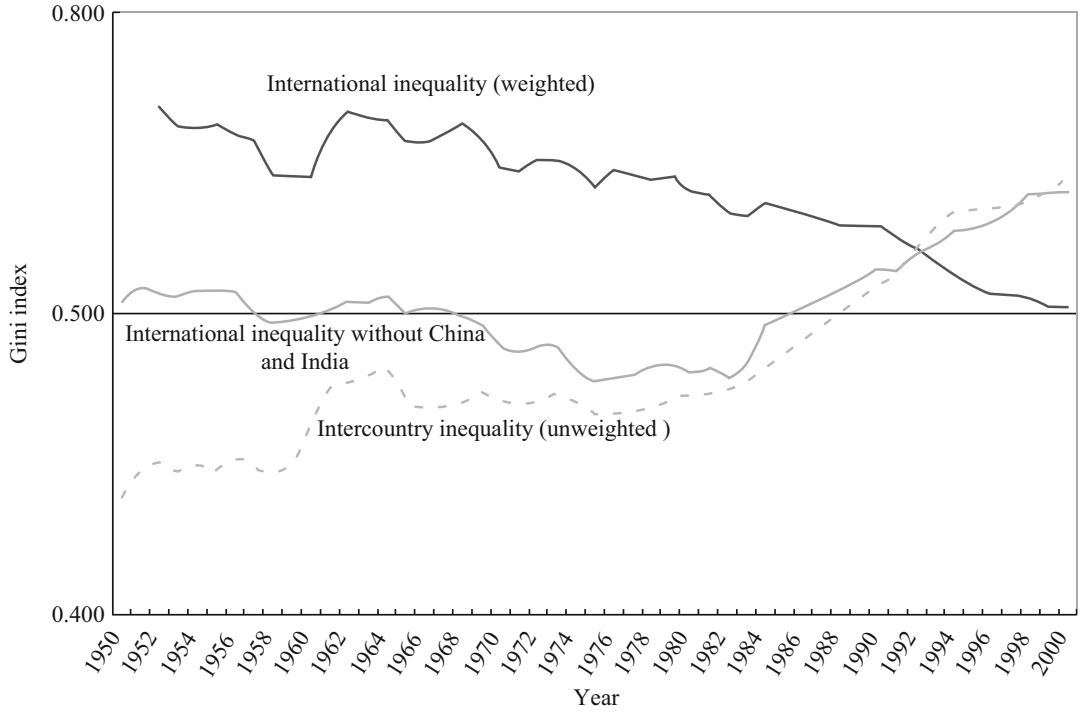
It is useful to distinguish three different concepts of world inequality (see Milanovic 2005; Bourguignon et al. 2004). The *inter-country* distribution measures the level of inequality across representative citizens of each country in the world. This is a distribution of unweighted gross national income (GNI) per capita. The *international* distribution uses country GNI per capita *weighted* by their population size: it measures the inequality in the distribution of the world's citizens if each citizen were assigned the average income of the country in which he or she resides, adjusted for purchasing power parity, instead of his or her own income.

Finally, there is the *global* distribution of individual incomes. This third concept lines up all citizens of the world (not countries) and calculates the distribution of their actual incomes, adjusted for purchasing power parity. *Global* inequality can be decomposed into inequality attributable to inequalities *within country* – that is, among persons within each country – and the differences of mean income *between countries*, that is *international* inequality.

The first section of this article describes the evolution of world inequality according to these various definitions and shows that they lead to somewhat contradictory conclusions, with inter-country inequality rising more or less continuously since the 1950s, international inequality declining, and global inequality increasing until the late 1980s and falling somewhat afterwards. The reason for these differences is easily understood, and has to do with population weights and the role played by giant developing countries like India or China and, to a lesser extent, with the evolution of inequality within countries. The second section tries to reconcile these various views on the evolution of world inequality by considering the mobility of world citizens within the income scale – this is similar to watching a movie rather than photographs of the world distribution of income at various points of time. The final section extends the income framework by providing some information on the evolution of world inequalities in a few non-income dimensions.

Evolution of World Inequality According to Alternative Definitions

The evolution of inter-country and international inequality between 1950 and 2000 is shown in Fig. 1, which shows that world income inequality, as measured by the Gini index, has been a story of increasing inter-country inequality and declining international inequality. (The Gini index is probably the most widely used measure of inequality. In theory it varies between 0 – perfect equality – and 1 – perfect inequality. Practically, it ranges from .20 to .25 in most egalitarian countries like the Nordic countries, and .6 or slightly more for the most inegalitarian countries in the world (for instance, Brazil or South Africa). Other measures are used below because of their decomposability property. But the evolution of the various definitions of world inequality is the same whatever the



Inequality Between Nations, Fig. 1 Inter-country and international distribution of income, 1950–2000 (*Source*: Milanovic 2005)

inequality measure being used.) A 20-year plateau was reached for inter-country inequality starting in the early 1960s, but the unequalizing trend resumed with the crisis of the world economy in the early 1980s. Due to differential demographic growth, no plateau is observed in the decline of the international inequality, but it can be seen that, since the 1987 or so, this decline is essentially fuelled by the fast growth of the two giant developing countries, namely, China and – to a lesser extent – India.

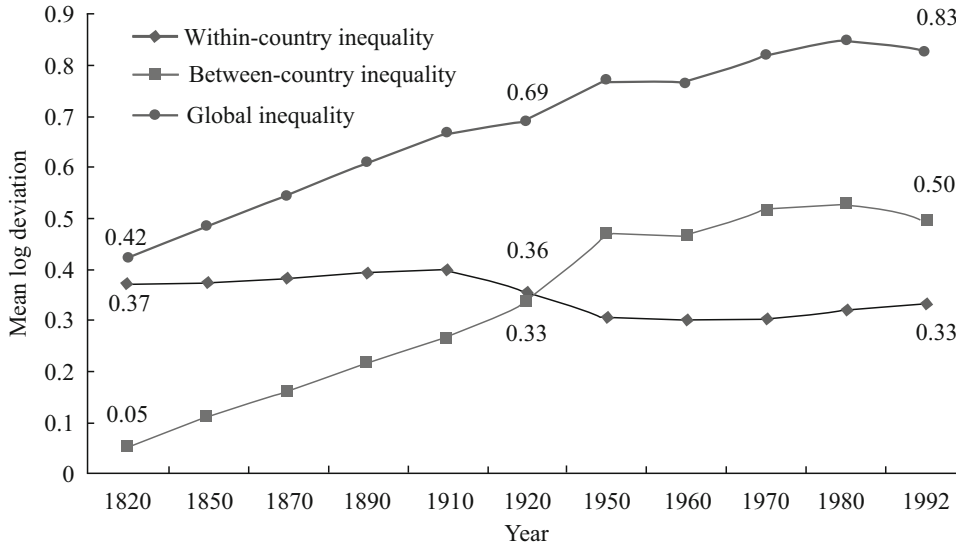
As China and India catch up to the world average, their equalizing effect on the international distribution of income will diminish. If they continue to develop at similar rates than in the past two decades, the effect of their growth will soon be unequalizing (Sala-i-Martin 2002c). Both inter-country and international inequality will then most likely increase unless countries at the bottom of the two distributions – sub-Saharan African economies in particular – begin to experience healthy growth. This suggests that, in the future, whether the world income distribution is equalizing or unequalizing will increasingly be a function of economic growth in Africa (and some other low-income countries), especially if population growth rates in Africa remain above world average.

To estimate *global* inequality, some knowledge of the distribution of inequality within each country is necessary. Producing estimates for long periods requires strong assumptions. Bourguignon and Morrison (2002) measure global inequality over 1820–1992 using historical estimates of countries' GDP per capita and rough figures on the distribution of income by deciles within countries to estimate the global distribution of income. They use income distribution information for a limited number of countries and assume that geographically and culturally similar countries or groups of countries have the same distributions. They then reconstitute the global distribution by assuming that individuals in each decile of a country's distribution have the same

income. With this method they estimate that the Gini coefficient for the global distribution increased from approximately 5 in 1820 to 66 in 1992, making today's world more probably more unequal than any single country. The mean logarithmic deviation shows the same dramatic evolution of global inequality in Fig. 2. But, because, this measure is decomposable into *within* and *between- country* inequality, it permits a better understanding of the forces behind that evolution. In particular, it turns out that *international* inequality was negligible at the turn of the 19th century (accounting for roughly 12 per cent of global inequality) but increased very rapidly until the Second World War. It then stabilized and started to decline after 1980. (The difference with Fig. 1 where international inequality declines more or less continuously after 1960 is due to the definition of countries – 33 'country groups' in Bourguignon and Morrisson as against 120 countries in Milanovic – and to the fact that Bourguignon and Morrisson considered discrete years rather than the whole annual series – for instance, 1960 is a 'low' point in the series shown in Fig. 1). Within country inequality, however, reached its peak around 1910 and declined dramatically (mainly due to equalizing forces in the now developed countries) between the two world wars, and started creeping back up only after the 1970s. The combined effect of these changes is an increase in the share of international inequality from roughly ten per cent in 1820 to more than 60 per cent by 1992.

For the recent past, trends in global inequality can be estimated using information from household surveys – which, for developing countries, have been available at regular time intervals only since the 1980s. Inequality within a country cannot increase drastically and indefinitely over the long term. Therefore, even though strong assumptions have to be made, Fig. 2, in which the evolution of the international inequality dominates, probably approximates rather well actual long- term trends. The same cannot be said for shorter periods. At the same time, estimates of global inequality become more uncertain over short time intervals because of problems of measurement and comparability between surveys for different countries or at different points of time.

What happened to global inequality in the recent past has been the subject of fierce debate in the context of globalization. In terms of method, Sala-i-Martin (2002a, b) uses an

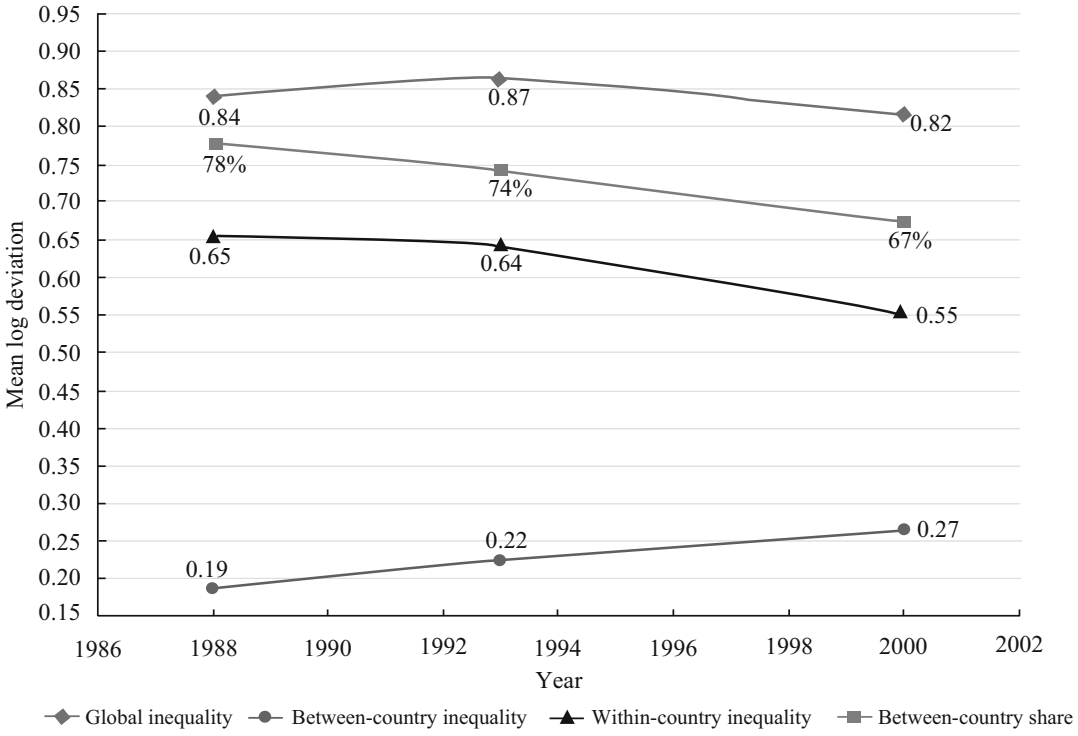


Inequality Between Nations, Fig. 2 Decomposition of global inequality into within-country and between- country inequality over a long period, 1820–1992 (*Sources*: Bourguignon and Morrison 2002; World Bank 2005)

approach that is similar to Bourguignon and Morrison (2002), combining GDP per capita figures and a constant rough continuous approximation of the distribution of income. Milanovic (2005) uses another method, estimating the parameters of country distributions on the basis of some predetermined functional form and grouped data from all available (comparable) household surveys over the 1980s and 1990s (which cover more than 90 per cent of the world population and world GNI). The sample of countries also differs non- marginally between the two studies.

In a three year comparison, Milanovic (2005) finds that global inequality increased slightly between 1988 and 1993 and then declined between 1993 and 1998. Sala-i-Martin (2002b, c) and others (see, for example, Bhalla 2002; Firebaugh and Goesling 2004) argue that it has declined. By and large, however, global inequality did not change much in 1988–2000. In all existing studies, variations of inequality do not exceed a few percentage points, whatever the inequality measure being used. This is in strong contrast with what was observed historically until the Second World War.

The latest and probably the most comprehensive estimate in terms of income distribution data being used is the World Bank (2005) – see Fig. 3. It confirms the evolution found by Milanovic. In agreement with Fig. 1, it also indicates that the share of global inequality that can be attributed to inequality between countries or *international* inequality declined steadily from 77 per cent around 1988 to 72 per cent around 1993 and to 67 per cent by around 2000. As global inequality stayed roughly the same during this time period, within-group inequality increased at a somewhat steady pace. These results are consistent with international inequality decreasing due to fast income growth in China and India – and with the



Inequality Between Nations, Fig. 3 Decomposition of global inequality into within-country inequality and between-country inequality using household survey data, 1980–2002 (*Source*: World Bank 2005)

evidence that inequality in China and in many other countries, including OECD countries, has been increasing over this period (see Ravallion and Chen 2007, for China; Atkinson and Brandolini 2004; Cornia 2004; Katz and Autor 1999; Gottschalk and Smeeding 1997). Interestingly enough, this evolution is the opposite of what was observed earlier in history. Does this evolution bring some support to the view derived from standard theoretical models of trade that globalization should tend to substitute inequality across countries by inequality within countries (see Bourguignon and Guesnerie 1999)?

What should we conclude from this review of evidence on the evolution of the world inequality of income? There is no doubt that the world has been increasingly unequal from the beginning of the industrial revolution until the end of the Second World War. But has it become more or less unequal afterwards? Should we rely on the *inter-country* definition and conclude for more inequality, or should we rely on the *international* – and *global* – definitions and conclude for less inequality? Should we give the same weight to China and the Liechtenstein, or should we have the world distribution of income depend on the relative performances of a few giant countries? It is argued in what follows that there is no right or wrong answer to these questions. Alternative definitions of world inequality correspond to different perspectives about the same evidence. The issue is to know whether this evidence is sufficient to form a final judgement or whether more is needed. The next section suggests that the common approach behind the figures reviewed so far may be misleading. Considering the income dynamics of countries, and country citizens, rather than viewing the 'anonymous' distribution at two points of time is more informative and somehow allows this apparent contradiction to be resolved.

Mobility on the World Income Scale

The approach of inequality in the preceding section focuses only on final outcomes and disregards initial starting positions. A better approach would be to track mobility. It departs from the conventional 'anonymous' view behind inequality measurement and can explain divergent opinions on changes in inequality between nations. If mobility itself forms part of the welfare criterion behind distributional judgements, then one is led to a conclusion about the change in the world distribution of income that is more nuanced and is consistent with both increasing and decreasing inequality. The main point is simply that the evolution of the distribution of income in the world since the 1980s has not been Pareto improving.

Some countries among the poorest and their inhabitants have lost income whereas the majority of world citizens have gained. Putting more emphasis on the former would then lead one to conclude that world welfare has fallen, that is, inequality has increased, whereas giving more weight to the latter leads to the opposite conclusion.

A simple way of tracking changes in the international distribution of income is to create 'mobility matrices' of world citizens moving over time from one income range to another (Bourguignon et al. 2004; Milanovic 2005). Such a mobility matrix is shown in Table 1. In the calculations behind this table, within- country inequality is

Inequality Between Nations, Table 1 Mobility matrix in absolute country per capita annual income levels (US dollars), 1980–2002

	Income in 2002				
Income in 1980	<710	711–1,100	1,101–2,890	2,890–10,000	10,001>
<710	1.28%	1.64%	0.00%	97.08%	0.00%
711–1,100	8.23%	3.89%	87.88%	0.00%	0.00%
1101–2,890	8.09%	0.56%	59.08%	32.28%	0.00%
2,890–10,000	0.00%	0.00%	0.98%	90.84%	8.17%
10,001>	0.00%	0.00%	0.00%	3.99%	96.01%

Source: Bourguignon et al. (2004)

assumed away and all citizens within a country are assumed to receive the same income. But there would be little difference if within country inequality were taken into account. In all cases, the world inhabitants that occupied the bottom range of the world income scale (less than 710 US dollars annually after correction for purchasing power parity (PPP), approximately the limit of the first quintile of the international distribution) in 1980 lived mostly in China and a few sub-Saharan African countries. In 2002, however, the Chinese had moved to an upper-income range whereas some sub-Saharan Africans who were initially in the second and third income range had fallen back to the first range. In other words, some poor people with income initially above 710 dollars in 1980 had fallen below that threshold by 2002.

Even though only eight per cent of each of the second and third income ranges fell into the bottom range over these two decades, this evolution clearly shows that no Pareto improvement has taken place in the world income distribution between 1980 and 2002. When considered in an anonymous way, as in the previous section, it may be the case that the average income of people in the poorest deciles of the international distribution of income has increased and may have even become closer to the mean world income. Yet this hides the fact that the composition of these deciles was very much modified. Chinese went out and were replaced by people from poor countries, initially richer than China, whose income fell after 1980.

Overall, whether the world distribution of income is judged as improving or worsening depends essentially on whether one considers that the increase in income of the Chinese and other poor people who climbed the income scale between 1980 and 2002 over- or under-compensates for the drop in the income of those people whose income fell. Looking at *international* inequality with population weights is equivalent to taking the former view, whereas focusing on *inter-country* inequality leads to the second conclusion. If the initial income position matters in assessing the social welfare of a population observed at a given point of time, then the social cost of falling incomes is not necessarily compensated for by the social gain of increasing incomes even if these changes take place towards and from the same income range. Deciding whether such a change increases or decreases social welfare requires some value judgements. The difference between those who feel that the world distribution of income has become worse and those who feel the opposite may simply be due to such differences in their value judgements.

Non-income Dimensions of World Inequalities

Most of the existing empirical evidence on inequality between nations concerns income, but recent studies have examined inequality in other dimensions of well- being, mainly health and education (see for example, Araujo et al. 2004; Deaton 2004; Goesling and Firebaugh 2004; Sala-i-Martin 2002c; Schady 2005). These studies indicate convergence in health and education indicators but divergence (or at least lack of convergence) in income. International inequalities in educational attainment and child mortality have been steadily declining, though improvements in life expectancy at birth have been set back since the early 1990s due to the devastating effects of HIV/AIDS and the difficult circumstances facing the former USSR and other transitions economies. Unlike global inequalities in income, global inequalities in educational attainments are attributable mostly to inequalities within countries.

What explains that there is convergence for health and education indicators and divergence for incomes? Deaton (2004, p. 109) points out that, while gains in income were undoubtedly important for improving nutrition and for funding better water and sanitation schemes, some countries made progress in reducing child mortality even in the absence of economic growth. These improvements came from the globalization of knowledge, facilitated by local political, economic and educational conditions. A possible explanation for the disconnect between the convergence in education and the divergence in

incomes is that education is not the only determinant of income and that the rise in per-worker schooling explains only a small part of the growth in output per worker.

Finally, it is worth insisting that there are major inequalities in voice and power between nations in participating to international decisions. These are discussed in detail in World Bank (2005). As Deaton (2004) puts it, poor countries lack the financial and human resources that would allow them to be equal participants in the international bodies in which decisions are taken that affect them and, beyond that, in setting the rules under which the international system operates.

Equity at the global level means that people should face the same opportunities for living the life they want regardless of where they are born. Income inequality among nations is only a sign that we are far from such a goal. Convergence in some non-income dimensions is an encouraging sign. Global action is possible in a number of areas to promote world equity, from improvements in international law and human rights, to promoting fairness in global markets, allowing free trade and free migration of labour, to more aid to the poorest, to a more equitable management of the environment and the global commons.

See Also

- ▶ Inequality (Global)
- ▶ Inequality (International Evidence)

Bibliography

Araujo, C., F. Ferreira, and N. Schady. 2004. *Is the world becoming more unequal? Changes in the world distribution of schooling*. Washington, DC: World Bank.

Atkinson, A.B.., and A. Brandolini. 2004. Global world inequality: Absolute, relative or intermediate? Paper presented at the 28th General Conference of the International Association for Research on Income and Wealth, Cork, 22 August.

Bhalla, S. 2002. *Imagine there is no country: Poverty, inequality and growth in the era of globalization*. Washington, DC: Institute of international Economics.

Bourguignon, F., and R. Guesnerie. 1999. L'économie mondialisée: inégalités entre nations hier, inégalité au sein des nations demain? Paper presented at an interdisciplinary seminar, Ecole des Haute Etudes en Sciences Sociales, Paris, 10 February.

Bourguignon, F., and C. Morrison. 2002. Inequality among world citizens: 1820–1992. *American Economic Review* 92: 727–744.

Bourguignon, F., V. Levin, and D. Rosenblatt. 2004. Declining economic inequality and economic divergence: Reviewing the evidence through different lenses. *Économie Internationale* 100: 13–25.

Cornia, G.A. 2004. Inequality, growth and poverty: An overview of changes over the last two decades. In *Inequality, growth and poverty in an era of liberalization and globalization*, ed. G.A. Cornia. Oxford: Oxford University Press.

Deaton, A. 2004. Health in an age of globalization. In *Brookings trade forum 2004*. Washington, DC: Brookings Institution.

Firebaugh, G., and B. Goesling. 2004. Accounting for the recent decline in global income inequality. *American Journal of Sociology* 110: 283–312.

Goesling, B., and G. Firebaugh. 2004. The trend in international health inequality. *Population and Development Review* 30: 131–146.

Gottschalk, P., and T. Smeeding. 1997. Cross-national comparisons of earnings and income inequality. *Journal of Economic Literature* 35: 633–687.

Katz, L.F., and D.H. Autor 1999. Changes in the wage structure and earnings inequality. In *Handbook of labor economics*, vol. 3A, ed. O. Ashenfelter, R. Layard, and D. Card. Amsterdam: North-Holland.

Kenny, C. 2005. Why are we worried about income? Nearly everything that matters is converging. *World Development* 33: 1–19.

Milanovic, B. 2005. *Worlds apart: Measuring international and global inequality*. Princeton: Princeton University Press.

Ravallion, M. 2004. Competing concepts of inequality in the globalization debate. In *Brookings trade forum 2004*. Washington, DC: Brookings Institution.

Ravallion, M., and S. Chen. 2007. China's (uneven) progress against poverty. *Journal of Development Economics* 82: 1–42.

Sala-i-Martin, X. 2002a. *The disturbing 'rise' of world income inequality*, Working paper no. 8904. Cambridge, MA: NBER.

Sala-i-Martin, X. 2002b. *The world distribution of income*, Working paper no. 8905. Cambridge, MA: NBER.

Sala-i-Martin, X. 2002c. Unhealthy people are poor people . . . and vice versa. Keynote address at the European Conference on Health Economics of the International Health Economics Organization, Paris, 7 July. Online. Available at http://www.columbia.edu/Bxs23/papers/parisconference.pdf. Accessed 24 June 2007.

Schady, N. 2005. *Changes in the global distribution of life expectancy and education*. Washington, DC: World Bank.

World Bank. 2005. *World development report 2006. Equity and development*. New York: Oxford University Press for the World Bank.

Inequality Between Persons

Anthony F. Shorrocks

Although inequality between persons can refer to a great variety of issues concerned with the disparate treatment and circumstances of individuals, economic discussion has focused on those aspects that relate to the acquisition and expenditure of income. As a consequence, the study of personal inequality has become largely synonymous with the distribution of income among individuals or households. Early contributors to this subject tended to provide an overall perspective on personal income distribution. In recent years, however, more attention has been paid to the particular dimension of inequality under investigation. Consideration has also been given to the precise way in which 'inequality' should be interpreted and measured, a trend most evident in the adjustments applied to observed incomes in order that the 'true' degree of inequality is revealed.

An initial distinction may be made between those studies which examine the origins of income dispersion and those which are interested in its consequences. The principal concern of the latter is inequality in living standards, or levels of well-being, and here the appropriate methodology is well established. For each household we require a measure of the level of its resources relative to its needs. The resource variable is typically identified with income, so that income distribution is the traditional point of departure in the study of unequal living standards. Ideally, however, income should be interpreted in a broad sense to include not only monetary receipts, but also unrealised capital gains, non-pecuniary benefits and household production which is not marketed. In addition, a long run income concept such as 'permanent income' or 'lifetime income' is preferred to the short run concept (weekly, monthly or annual) which applies to most of the readily available data. These incomes should then be adjusted to allow for different household circumstances. One type of adjustment concerns family characteristics, such as the number and ages of family members, and is accomplished by the use of household equivalence scales. A second type of adjustment relates to the environment in which the household operates and covers such factors as the prevailing level of commodity prices; the shelter, heating and transportation requirements associated with household location; and the level of provision of public goods and services. The aim, as before, is to achieve comparability between households in different circumstances. Needless to say, while this programme of adjustments may be generally accepted as the ideal, most empirical studies of living standards fall a long way short of the target.

A much larger body of literature is concerned with the causes of income inequality. This work focuses on the experience of individuals in factor markets and covers a wide range of issues on which opinions seldom agree. It will be helpful to begin by splitting the income y of an individual into components and writing

$$y = y_1 + y_2 + \ldots + y_n \tag{1}$$

$$y = r_1 x_1 + r_2 x_2 + \ldots + r_n x_n, \tag{2}$$

where r_i and x_i are the 'price' and 'quantity' associated with the ith component of income. A decomposition of this form would be appropriate if the x_i denoted the individual's endowments of productive factors, such as labour, capital and land, and the r_i represented the corresponding factor prices. This immediately suggests two principal causes of income inequality: differences in the endowments of productive resources which individuals own; and the structure of factor prices determined by the combination of institutional, market and social forces which we will call the *common environment* of individuals. Further refinement of this line of reasoning can be achieved by extending the coverage of the x_i to include a variety of other characteristics, and by looking back towards the source of the characteristics: to inheritance, in the form of genetic traits, material wealth and family advantage; to innate and acquired skills; and to the choices individuals make in respect of occupation, location and

workhours. There is also, inevitably, a portion of income, often attributed to 'chance' or 'luck', which is not systematically related to any of these factors.

Theories of income distribution differ not only in the particular influences and mechanisms that are stressed, but also in their view of what a theory of income distribution should set out to accomplish. One aim is to account for the overall degree of inequality at any date, and the pattern of changes in aggregate inequality that take place over time. Another topic of interest is the characteristic shape of the frequency distribution which incomes tend to follow. A third objective is to explain why different individuals happen to have different incomes. All three of these issues are valid concerns, and all would be addressed in a satisfactory general theory. On the whole, however, a distinct literature has developed on each of the questions. These are reviewed in turn below.

Changes in Income Dispersion Over Time

Explanations of changes in income inequality over time have typically drawn attention to the features of the common environment and to the consequent pattern of factor prices. Factor prices play a significant role in most of the discussion of income distribution prior to 1900, and even up to the middle of this century, as indicated by the selection of papers published by the American Economic Association in 1946. This is perhaps a reflection of the rigid social structure in the 19th century, which made it natural to assume that the resource endowments of individuals remained relatively constant over time. In those circumstances the first priority was to account for the level of factor prices or factor shares. Once this was done, aggregate factor payments could be allocated among individuals according to a prearranged pattern of entitlement. A theory of factor prices was, in effect, a sufficient explanation for the distribution of personal income.

The tendency to submerge the theory of personal income distribution within the grander themes of Labour, Capital and Land was not without its critics. Cannan (1905) was prompted to suggest that a student seeking an explanation for the riches and poverty surrounding him would return home in disgust from a typical lecture on the subject. He argued that more attention should be paid to the way that aggregate factor payments were shared between individuals, a suggestion taken up with enthusiasm by Dalton (1920) in one of the earliest and most outstanding volumes devoted to personal income distribution. The ideas of Cannan and Dalton had little immediate impact. But the importance they attributed to inheritance has certainly been echoed in subsequent research, most notably in connection with the sources of wealth inequality.

More recently, the explanation of long-run trends in income dispersion has been particularly associated with the work of Kuznets and Tinbergen, and again regards the common environment as the ultimate source of change. The programme of research that has developed from Kuznets (1955) is concerned with the relationship between economic growth and the distribution of income within countries, and sees demographic movements as a major influence on inequality. Countries begin with a fairly homogeneous population, largely employed in the traditional sector. In the course of development, individuals transfer into the modern sector causing income inequality to first rise and then fall, as the modern sector becomes dominant. Inequality within both the traditional and modern sectors may, however, remain constant. This suggests that observed variations in inequality could be spurious, reflecting the way in which data is recorded rather than a real change in the relative income positions of individuals. Other demographic factors, such as shifts in the age structure and household composition have also been cited as having a similar impact.

Tinbergen (1975) appeals to the common environment influences that determine the relative earnings of skilled and unskilled workers. In his view, the observed trend in income inequality is the outcome of a race between technology and education. Technological progress creates additional demand for skilled workers, while improved educational opportunities increases the supply. The direction of movement of the skilled–unskilled wage differential depends on

the relative strength of these two forces. Over the course of this century, education has advanced faster than technology, driving down the relative earnings in professional and skilled occupations with notable consequences for income dispersion. The distinguishing characteristic of Tinbergen's argument is the attention given to the operation of factor markets. Many other studies have been concerned with the role of education and training, but they tend to emphasize the process of skill acquisition and treat factor prices as exogenous data.

The Pattern of Income Frequencies

It has long been recognized that the density function for incomes has a characteristic shape, sufficiently regular and well documented to merit special attention. The major early contribution to this line of enquiry was undoubtedly Pareto, who, in a series of publications, including his *Cours d'économie politique* (1896), assembled evidence on personal incomes spanning more than four centuries and expounded his universal law of income distribution. Pareto noted that the data were closely matched by the formula

$$\text{In } N = \text{In } A - \alpha \text{ In } y, \tag{3}$$

where y is a given level of income and N is the number of people with incomes above y. Furthermore, the slope coefficient α was always approximately 1.5. This, he argued, could not be a coincidence. The statistical regularity must indicate a natural state of affairs which would tend to reassert itself if, for any reason, the income distribution departed temporarily from its stable equilibrium.

Pareto's results, and the inefficacy of redistributive policies which they seemed to imply, soon attracted both dedicated support and hostile opposition. The increasing availability of data eventually undermined the strong version of Pareto's law. However the tendency for the upper tail of incomes to follow the Pareto curve (3) is well established, and remains one of the 'stylized facts' concerning the distribution of both income and wealth. Pareto also had a profound influence on many of the methodological developments that have subsequently taken place. His interest in the collection and summary of data, the parametric description of income frequencies, and the identification and investigation of observed statistical regularities, are all strongly echoed in later research.

Although high incomes tend to follow the Pareto relationship the Lognormal distribution is a better representation over the whole income range. This provides a clue as to how the observed pattern of incomes could arise. For just as normal distributions result from a large number of small and statistically independent effects which combine additively, so lognormal distributions emerge when the effects combine multiplicatively. More formally, if we replace income in (1) with the logarithm of income to obtain

$$\text{In } y = y_1 + y_2 + \ldots + y_n, \tag{4}$$

and assume, say, that the y_i are identically and independently distributed variables, then the income pattern will be approximately lognormal when n is sufficiently large. In this argument it is the statistical properties, rather than the origins, of the components y_i that are significant. They may, therefore, be treated as unspecified random effects.

The notion that income distribution can be viewed as the outcome of a process governed by a large number of small random influences was developed by a number of authors, most notably Gibrat (1931) who formulated his 'law of proportionate effect', and Champernowne (1953) who demonstrated how the Pareto upper tail could emerge as a feature of the equilibrium distribution of a Markov Chain. Further modifications to these models were later shown to be capable of generating, as the limit of a stochastic process, a variety of other functional forms which have features in common with the lognormal and Pareto distributions, and which may be regarded as reasonable descriptions of the frequency distribution of income. As explanations of income inequality they have been criticized on the grounds that chance or luck, rather than systematic personal

or market forces, appears to be the principal determinant of income differences. But the significance attached to this complaint depends on the question that is being addressed. If we are primarily interested in accounting for the overall features of the density function of incomes, it may be appropriate from an aggregate perspective to treat the incidence of personal success and failure as a random event, while simultaneously accepting that success and failure may be rationalized at the level of specific individuals.

Income Differentials

The dominant theme of recent research on personal income inequality is the explanation of income differentials – the reasons why particular individuals have different incomes. This work has several distinctive features. One is an increasing concern with empirical issues, and with the empirical evaluation of competing theories, facilitated by the availability of large-scale survey data and the means to process the information. Nowadays the question is not so often which factors have an impact on income distribution, but which factors have the most quantitative significance as explanations of observed income differences. Another distinctive feature is the emphasis placed on the distribution of earnings, again partly a result of the quantity and quality of earnings data. Earnings inequality is important, not only because wages and salaries form such a large proportion of income, but also because it reflects the extent to which labour markets operate fairly. As a consequence the explanation of the inequality of pay has become the main battleground for opposing views on the origins of personal inequality.

One method of approaching the question of earnings differentials is to regard the labour market as being composed of a set of individuals with different personal traits P, and a set of job opportunities offering various combinations of characteristics J. The process of matching people to jobs then generates a relation between personal traits, job characteristics and pay which is typically captured in an earnings equation of the form

$$\ln\ w = \alpha_1 P + \alpha_2 J + R, \tag{5}$$

where w denotes earnings or hourly wage rates, and R is a non-systematic or random effect. In this formulation, the coefficients α_1 and α_2 may be interpreted as the 'prices' which the market imputes to the various characteristics. Their values are often estimated from empirical data. Notice that equation (5) has similarities with (2) and, more especially, with (4). This indicates that the separate influences combine multiplicatively, rather than additively, as Gibrat had suggested earlier.

Many different views on the determinants of earnings can be accommodated within equation (5). One common argument claims that earnings are related to the ability or productivity embodied in individuals. Certain personal traits may therefore be valued because they indicate the actual productive performance that derives from either natural ability or the skills acquired as a result of education, training and work experience. Furthermore since perceived, rather than actual, productivity is rewarded in the market place, other characteristics may also be valued if firms believe them to be correlated with relevant variables that cannot be observed. This can account for the 'prices' imputed to educational credentials, family background, gender and race. There are, however, alternative explanations for sex and race discrimination: consumer and employer prejudice towards the goods and services provided by disadvantaged groups; and exploitation by firms of the different supply elasticities that arise from personal circumstances and role specialization within the family.

The structure of job opportunities and the prices imputed to job characteristics lead to a different set of considerations. One line of argument, associated with segmented labour market models and the Job Competition model (Thurow 1976), stresses the significance of the distribution of jobs across occupations and industries, which depends on the state of technology, the structure of markets and the other social and institutional factors contained in the common environment. It is this job distribution, together with customary wage and salary differentials, which is the

principal determinant of earnings inequality. Personal characteristics appear to be important only because they are used by firms to ration entry into the more attractive jobs.

This argument suggests that the desirability of any given occupation is directly related to its imputed price. Exactly the opposite conclusion applies if the prices of the characteristics are interpreted in terms of Adam Smith's (1776, Book 1, ch. X) concept of compensating differentials. For if individuals can choose to trade-off income against job characteristics such as occupation, location and working conditions, it is precisely those jobs with the least desirable features that need to pay more in order to attract an adequate workforce.

The notion that some part of earnings may compensate for other job features, and that the process of choice may help to explain observed income variations, has special significance in the analysis of personal inequality. This may be seen by considering a situation in which people are faced with a set of employment prospects each of which offers a level of income and a combination of other characteristics. If all individuals select from the same set of options, there is clearly no 'true' inequality in the sense of unequal opportunity. Yet people with different tastes will choose different alternatives, so observed incomes will typically vary. It follows that observed income dispersion may well exaggerate the true degree of inequality if some income differences are attributable to choice.

Individuals do not, of course, all face the same set of options. So different opportunities, as well as different choices, contribute to observed inequality. Much of the discussion of the determinants of earnings can be viewed in the context of the distinction between these two factors. Indeed, the most controversial aspects of the study of income distribution often reflect conflicting opinions on the relative importance of the 'true' component of inequality arising from different opportunities and the 'spurious' element of inequality that results from choice. Notice that choice is not the only mechanism that separates opportunities from outcomes. Chance also has a role to play if uncertain prospects are among the set of available options. It is therefore most appropriate to decompose inequality into three components: choice, chance and unequal opportunity. The influence of chance will, however, tend to disappear when we examine the typical experience of a group of individuals.

Those who stress the importance of unequal opportunity tend to focus attention on the contribution of natural ability, family background and discrimination. Here we might, perhaps, distinguish between two aspects of unequal opportunity: the unequal inherited endowments associated with natural ability and family background; and the unequal market treatment associated with discrimination. In contrast, the impact of choice is most clearly seen in the decisions relating to hours of work and geographical location. Individual preferences can also explain the choice of occupation and length of training. Thus, for example, a naive version of the Human Capital model suggests that the level of acquired ability is freely chosen under conditions of equal opportunity, so that the earnings differentials corresponding to education and training are purely compensatory. Refinements of the model, however, allow training opportunities to be influenced by ability and family background, and these factors, together with discrimination, are important elements of those arguments which emphasize the lack of equal and open access to training programmes. The impact of choice on skill levels depends, therefore, on the precise process by which skills are acquired and augmented.

The debate on the relative importance of unequal opportunity and choice for earnings inequality has its counterpart in the study of investment income, via the determinants of wealth distribution. Here, individual preferences are captured in the motives for saving: a desire to provide for retirement, to make bequests, or simply to practice thrift. Choices based on these preferences then determine savings behaviour and can account for some wealth differences in terms of past accumulation. Unequal opportunity, on the other hand, appears principally in the guise of material inheritance, but may also arise from the differences in incomes and family circumstances that affect the opportunities for saving.

See Also

- ▶ Discrimination
- ▶ Labour Market Discrimination

Bibliography

American Economic Association. 1946. *Readings in the theory of income distribution*. Philadelphia: Blakiston.
Atkinson, A.B.. 1983. *The economics of inequality*, 2nd ed. Oxford: Clarendon Press.
Cannan, E. 1905. The division of income. *Quarterly Journal of Economics* 19: 341–369.
Champernowne, D.G. 1953. A model of income distribution. *Economic Journal* 63: 318–351.
Dalton, H. 1920. *The inequality of incomes*. London: Routledge.
Friedman, M. 1953. Choice, chance, and the personal distribution of income. *Journal of Political Economy* 61: 277–290.
Gibrat, R. 1931. *Les inégalités économiques*. Paris: Sirey.
Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45: 1–28.
Mincer, J. 1958. Investment in human capital and personal income distribution. *Journal of Political Economy* 66: 281–302.
Pareto, V. 1896. *Cours d'économie politique*. Lausanne: F. Rouge.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E.R.A. Seligman. London: J.M. Dent, 1910.
Thurow, L. 1976. *Generating inequality.* London: Macmillan.
Tinbergen, J. 1975. *Income distribution, analysis and policies*. Amsterdam: North-Holland.

Inequality Between the Sexes

E. Boserup

Economic theory concerning inequality between the sexes focuses upon inequality in wages, job recruitment, promotion and dismissal, for women and men with similar qualifications and availability. Neoclassical theory explains these inequalities as a result of free and rational choice, based upon the biological differences between the sexes. According to Becker (1981), women's role in reproduction makes it rational for women to specialize more in family skills, and men more in labour market skills, and parents make a rational choice for their children by preparing them for different careers. When women's reproductive role is reduced due to the decline of birth rates, women's availability for the labour market increases, and they begin to invest more in labour market skills than is the case in countries with continued high fertility. So sex-related differences in level and types of human investment and availability provide the explanation for the differences in wages, types of work and promotion.

By focusing upon the biological differences between the sexes, neoclassical theory selects the features which distinguish inequality between sexes from inequality between other discriminated groups, that is the young versus the old, or foreigners versus members of the dominant ethnic or national group. All these inequalities have been characteristic features of human societies since prehistoric times. The basic principle in the organization of societies is that only members of the superior group have adult status or civic rights, while the members of the inferior groups depend upon the benevolence of the 'adults'. In most societies, economic and social development have reduced the inequalities, but nowhere have they been completely eliminated, and the traditional power of the superior male group over the inferior female group cannot be ignored in the economic analysis of inequality between sexes. The power of the male group over the female one is supported by access to the best technology and a monopoly in learning how to use it (Boserup 1970). Men's monopoly in the use of weapons, superior hunting equipment, and animal-drawn agricultural equipment, is of ancient origin. But even in societies where men have shifted to tractors and other industrial inputs, women often continue to use primitive hand tools for the operations assigned to them, and even in modern mechanized industries, men distribute the tasks and assign the unskilled, routine operations to the female workers.

In primitive subsistence economies, woman's reproductive role does not prevent her being assigned the most onerous tasks with incessant daily toil, and if the mother's work prevents her from taking care of young children, these are

cared for by older sisters or other members of the group. At a later stage of development, when specialization of labour leads to the transfer of an increasing share of the labour power of the family to outside work, the reproductive role of women contributes to explaining why more women than men continue to work in the family and for the family, either as unpaid family members or as domestic servants. However, due to their superior status, men have the right to dispose of money incomes earned by female family members within or outside the family enterprise. There may be regional and local differences in women's status, but in most traditional societies women cannot dispose of money or undertake monetary transactions, accept employment or move away from the locality where they live and work, without the permission of a male guardian who decides all these matters, as well as family matters, like marriage, divorce and the fate of the children. The right to take part in decisions on public matters is reserved for members of the male sex.

Gradually, as technological development transfers an increasing number of products and services from family production to production in specialized enterprises and institutions private or public, there is no need for the full labour power of all female family members in the household. Through the same process the family economy becomes more and more dependent upon money income to purchase the products and services which the family no longer produces, and to pay the taxes which finance the growing public sector. As a result, increasing numbers of women become money earners. At this stage, women's ability to engage independently in economic and other transactions, and their lack of responsibility, becomes a handicap not only to themselves but also to their employers, creditors, customers and guardians, as well as to public authorities and male family members who must support them, if they are unable to support themselves and their dependents because of economic disabilities.

In some European countries, 'market women', who were often middle-aged women with dependents, attained adult status many centuries ago. Later, when it became customary for young girls to work for wages before their marriage, and for other single, divorced and widowed women to support themselves and their dependents by wage labour, or by self employment, these categories of women were granted 'adult' status in economic affairs; but married women continued to be denied adult status. In most industrialized countries, married women first attained adult status when further reduction of the domestic sector, together with the decline of birth rates, radically increased their participation in the labour market and made their work in the labour market an important part of the national economy. In most developing countries, women, whether married or not, are still denied adult status in economic affairs; in some countries it severely limits their labour market participation, in other cases it limits the business activities they are able to accomplish.

Human capital investment in 'market skills' becomes more and more important with economic and social development, while investment in family skills loses in importance when more and more activities are transferred from the family setting to private enterprises or public institutions. When the responsibility for physical protection is transferred from the family to the government, and formal education is introduced, educational level may replace the ability to use weapons as a status symbol for male youth. The priority given to boys over girls in formal education is not only a result of their larger labour market participation, as suggested by Becker, but also a means to preserve a higher male status, by letting men reach higher educational levels than women.

The status of parents may require that their daughters be educated as well, but that boys should not lose status by receiving less schooling than their sisters, while to preserve the superior status of the husband, the wife must not be more educated than he is. Universities were long closed to women, and in many countries the difficulties of obtaining marriage partners for educated women make both parents and daughters afraid of continuing their education. The low marriage age for girls, another means to preserve male status in the family, may also prevent continuation of the education of girls. The differences between the sexes in educational levels serve to reinforce inequality not only in the family, but also in the labour market. With

economic development the difference becomes limited to the highest educational levels but it has not disappeared, even in countries with very high and uninterrupted female labour force participation.

Usually, differences in access to technical training for girls are much larger than differences in access to formal education. From the day women began to work for wages in urban activities, men have insisted on their priority right to skilled, supervisory, and other better paid work. Both in guilds, and later in industries and public service, men became apprentices and skilled workers while women remained assistants to the male workers, unskilled or semiskilled, working under male supervision. In most cases, male trade unions continued the fight of the guild members against rights for women to training, and even to membership of the organization and right to work in the trade. The inferior position of women was defended by the short stay in the labour market of young girls before they married, with no account taken of the large number of spinsters, poor married women and female heads of households, who were permanent members of the labour force both in European and in many non-European countries.

In addition to the lower position of women in the job hierarchy, female wage rates are usually much lower than male wage rates for similar work. Only in periods of great shortage of labour, for instance in wartime or in agricultural peak seasons, may female wages temporarily rise to the level of male wages. The fact that these wage differences are related to sex, and not to the burden of dependency, belies the usual explanation for them. They are a result of the principle of male superiority, and neoclassical theory has helped to make the principle acceptable. Since the theory assumes that differentials in wages equal differentials in marginal productivity of labour, the lower wage rates for women could be taken as a confirmation of the general assumption of female inferiority, which also applied to women as workers.

The superior status of men is supported when women doing similar work get lower wages; when a wife is prevented from earning as much as her husband, he preserves his superior status as principal breadwinner, even if he is too poor to enjoy the even higher status of being the only breadwinner in the family. Training girls in low-wage occupations and discriminating against women in recruitment for 'on the job training' or access to 'learning by doing' supervisory work, reduces the risk that male staff will lose status by being supervised by women.

When employers in private enterprises and public service pay males higher wages than females for similar work, they include the higher male wages in their production costs, even if that reduces the demand for products made primarily by male labour. If an enterprise or a trade has difficulties in competing, due to the payment of high male wages, employers will not reduce the wage differential, but will instead try to get the workers and the trade unions to accept the recruitment, or additional recruitment, of women. If they succeed, the trade will become less attractive to men, and the labour force will gradually become female as has happened to many trades in which trade unions were weak. The separation of the labour market into masculine and feminine trades and jobs becomes even more pronounced if the principle of equal pay for equal work is introduced by law or labour contract, since sex specialization makes it more difficult to prove that the work paid at different rates is 'equal'.

Inequalities between men and women in the labour market and in the family reinforce each other. While Becker assumes a harmony of interests between the marriage partners and an equal distribution of consumption and leisure between them, Sen (1985) uses bargaining theory to explain the observed inequalities in consumption and leisure, which in some countries include differences in coverage of calorie requirements and in access to health care between husband and wife, and between boys and girls. The wife's bargaining position is directly related to her access to the labour market and position in it, but her bargaining position is also weakened because women are likely to perceive inequalities as natural, and make no objections against them. This feature is due to the family socialization of girls from a young age. In many societies, girls are taught that they are less valuable human beings than their brothers, and virtually everywhere girls

must help their mothers to provide domestic and personal services for their brothers, who are allowed much more freedom and leisure.

Even in countries with high and perpetual labour force participation by women, girls' education and training within the family focus on child care and domestic activities, and on beautifying themselves to be able to make a good match and reduce the risk of divorce and abandonment, while boys' interests are stimulated in all other fields. Usually, girls are taught to be obedient, to be modest and to do routine jobs without protest, while boys are encouraged to be enterprising, even aggressive, and more self-confident. The inferiority feelings of the girls may induce them to invest less in education and training than boys, as suggested by Arrow (1973), but even if they have the same formal education and training as male competitors, women are likely to lose in competition with males in the labour market. Girls, who are socialized to accept routine jobs and to be modest and obedient, are unlikely to demand good jobs and advancement, or in other ways to fight actively for their interests in the labour market, even when there are few prejudices against them. Much female aptitude for routine and precision work, unsuitability for leadership and unwillingness to take responsibility results from family socialization in the first years of life. Most often, the schools continue in the same vein, but even when schools aim at abolishing inequality between the sexes, the teachers may be powerless, due to family socialization of pupils of both sexes.

In industrialized countries, the last few decades have seen an acceleration of related and mutually reinforcing changes in technology, labour participation by married women with small children, and birth rates. Decline of birth rates to below replacement level, and increasing female labour force participation provide an inducement to the improvement of household technologies, and the introduction of new products and services as substitutes for women's traditional activities and child care. These technological and social changes further induce increasing female labour force participation. A rapidly increasing proportion of married women continue their money-earning activities without reducing work hours during the period when they have small children. But the traditional sex hierarchy is dying very slowly, and although birth rates are low, female levels of education and professional training fairly high, and labour market participation high and continuous, reductions in sex differentials in earnings have been moderate, if any. Earnings in female occupations, including those requiring professional training, are lower than in male occupations with similar requirements. Except for a small female elite, women continue to occupy the positions at the bottom of the labour market within each occupation, as assistants to men, and often supervised by men even in otherwise female occupations.

Married women with full-time work and young children have much longer working hours than men and little leisure because male patterns of work have changed very little, in spite of reduced working hours and the increasing amount of money wives contribute to family expenditure, and also because of the lack of child-care facilities in many countries. However, in spite of the differences between male and female earnings, most women in the industrialized countries have become less dependent upon male support because of the general increase of all wages and the reduction of working hours in the labour market. Therefore, women can support themselves by work in the labour market, if they choose to, and with the aid of obligatory contributions from the father, and public support to female-headed households, they can support children, although the living standards of female-headed households are usually much lower than those of male-headed households. Consequently, many young women react against unequal work burdens by demanding divorce or leaving the home, or by not entering into a formal marriage or cohabitation. Others react by reducing birth rates even further. Contrary to earlier patterns, female applications for divorce have become more numerous than male ones in some industrialized countries. These social and demographic changes serve to make young men, and public opinion in general, more inclined to consider women's demands for more equality.

I

In many developing countries, economic and social development are producing changes in female labour force participation and birth rates which resemble earlier changes in industrialized countries. Family legislation has been modernized, there is legal equality or less legal inequality between the sexes, access to divorce has become less easy for men and easier for women, and better access to the labour market provides women with some possibilities for self-support in case of divorce and widowhood. Age differences between the spouses are declining due to higher female marriage age, birth rates are declining, and women's position is gradually improving.

But in many other developing countries, either economic changes are few, or male resistance to changes in the traditional status of women is strong. Except for voting rights to parliaments with little influence, women continue to be legally minor, and in many cases their situation has deteriorated because technological changes, or changes in land tenure, have deprived them of traditional means of self-support. In some countries, the labour market continues to be closed not only to married women, but also to deserted women, divorcees and widows, and if labour market shortages occur, they are met by large scale imports of male labour. In these countries birth rates remain high in spite of economic development. For women, economic support from sons is the only alternative to destitution, when the husband dies or ceases to support his wife, and women also desire to have many sons as a means to reduce the risk of abandonment and divorce.

See Also

- ▶ Discrimination
- ▶ Gender
- ▶ Labour Market Discrimination
- ▶ Women's Wages

Bibliography

Arrow, K. 1973. The theory of discrimination. In *Discrimination in labour markets*, ed. O. Ashenfelter and A. Rees. Princeton: Princeton University Press.

Becker, G. 1981. *A treatise on the family*. Cambridge, MA: Harvard University Press.

Boserup, E. 1970. *Woman's role in economic development*. London: Allen & Unwin.

Buvinic, M., M.A. Lycette, and W.P. McGreevey. 1983. *Women and poverty in the third world*. Baltimore: Johns Hopkins Press.

Cain, G. January, 1985. Welfare economies of policies towards women. *Journal of Labour Economics* 3(1), special issue, 375–96.

Sen, A. 1985. *Women, technology and sexual divisions*. New York: United Nations.

Tilly, L., and J. Scott. 1978. *Women, work and family*. New York: Holt, Rinehart & Winston.

Inequality of Pay

Henry Phelps Brown

The difference between the hourly rates of pay for two jobs might be seen in the same way as that between the prices of a ton of copper and a ton of steel: there is a common unit of quantity, but the conditions of supply and demand for the two articles are largely independent, and the difference between their prices is only an arithmetic by-product. But people do attach significance to the differences between rates of pay, and their ideas about what these should be help to fix particular rates. The relations between the rates of pay for different jobs are termed *differentials* when the jobs compared lie at different grades within the same occupation or industry, and *relativities* when they are in different ones. Both sorts of comparison are possible because jobs are defined by their requirements – what physical and mental ability, length of training, experience, tolerance of adverse working conditions, and the like, they demand from anyone who is to do them adequately; and these requirements are regarded as being common to jobs of all kinds, but present in different amounts and proportions. The requirements of given jobs are assessed intuitively by those who make practical judgements about the fairness of differentials and relativities, and are set out explicitly in the procedure known as job

evaluation. The rate for a job is regarded as made up of the shadow prices of the capacities to meet the several job requirements, together with the extent of each requirement in the given job.

The question then arises, how those shadow prices are determined: how comes it about that skill commands a higher price than muscle? Two answers have been given – convention, and supply and demand. The case for convention opens with an appeal to everyday knowledge of how people insist on the maintenance of customary relations. They require that relative pay conform with status. That the labourer's rate stood at two-thirds of the craftsman's in the building industry of Southern England over more than six centuries (Phelps Brown and Hopkins 1981) can be attributed only to convention. If women's rates had been simply proportioned to productivity, it is hard to account for their relative rise when the rule of 'equal pay for equal work' was enforced in the UK and the Netherlands. Here convention had been keeping pay down, but John Stuart Mill remarked long ago (1848, II, xiv, 2) that it was keeping the relative pay of clerks up, after the increase of supply of clerical capacity through extended education had tended to lower it. These observations are all consistent with differences of pay being set to match accepted gradations of status. But on the other hand there are all the instances of market forces moving a particular rate, and so changing a relativity, when there has been no question of an antecedent change in status. Differentials that have long remained constant have changed when an upheaval has loosened the hold of custom, and the way in which they have changed can be explained by market forces. The Soviet-type economies use the differentials and relativities in their pay structure as incentives to attract and deflect the supply of labour to particular employments. In the West, competition of employers for a new skill such as computer technology opens up a differential over the pay commanded by other qualifications at the same level. A mental experiment indicates that if those who need a long and costly training before they can meet a job's requirements – say surgeons – were to be paid no more than the unskilled, then though a certain number would still enter the profession out of interest in the work or a sense of social responsibility, it would not be possible to maintain the numbers for which consumers have shown themselves willing to pay at actual rates.

We conclude that the basic reason for the inequality of pay is that, on the side of demand, users are willing to pay different amounts for the capacity to meet the requirements of different jobs; and on the side of supply, that unless a certain rate of pay is provided, labour capable of doing the job will not eventually be forthcoming in the amount that users wish to employ at that rate. If status and pay commonly agree, that is because the personal capacity that confers status also commands a higher price because of its productivity. But there are zones of tolerance within which market forces do not fix rates closely, and here convention may prevail.

On this view, the major obstacle to the reduction of inequality lies in the limitations of the supply of labour to the better-paid jobs. These are to be found in the genetic distribution of personal potential, and next in the distribution of the factors moulding capacity in early childhood, from the homes that foster development to those that thwart it. The quality and extent of education are then limited by the availability of institutions, and the cost of maintenance of students that falls on their families. Similar limitations restrict the numbers obtaining training following education. Readier access to education and training should lower the supply price of labour to the jobs with more exacting requirements and higher pay; but the limitations imposed by heredity and early upbringing will remain (Rutter and Madge 1976; Phelps Brown 1977, chs. 6, 7 and 9).

So far the discussion has concerned the pay of different occupations, but the differences between the earnings of individuals in the same occupation also demand consideration. These are commonly wide. Some of the range arises from short-period fluctuations in bonus and overtime. Earnings also vary with age, and the inequality of lifetime pay is much less than that of pay at any one time. But a great part of the range is due to differences of individual performance. There are also some systematic forms of differentiation. Regional

differences in the rate of pay for the same work are often substantial. In local labour markets quite large differences are found between the pay for a given occupation in different firms. It is generally higher, the bigger the firm. There may be discimination against ethnic minorities or women. Discrimination 'before the market' occurs when the victims are denied equal access to the means of acquiring capacity. Discrimination 'within the market' occurs when some persons receive lower pay than others by reason of their ascriptive characteristics and not because of, or in proportion to, their lower capacity. It appears likely that discrimination 'within the market' is much less considerable than discrimination 'before the market'.

The combined outcome of differences of pay between jobs and among individuals in the same job is a distribution of individual earnings. The form of this distribution confronts us with a striking social regularity. When Lydall (1968) brought together comparable data of earnings from more than thirty countries, he found a common form of distribution. This was unimodal, with a long upper tail. The central part was closer to a log-normal than a normal distribution, but both tails were thicker or longer than in the log-normal, and in particular the upper tail was fitted closely by Pareto's formula. Generally, distributions of earnings are now taken to be log-normal but with a Paretean upper tail. Soviet planners are understood to treat the distribution of earnings as log-normal.

The common form does not imply an equal measure of inequality, though there are some striking instances of this. The distribution of the earnings of manual men found in the British wage census of 1886 agrees closely with that found in the 1970s. Bergson (1984) found 'a rather striking similarity in equality, as measured, between the USSR and Western countries,' though the USSR distribution lacks the Paretean tail. But differences appear among European countries. 'Britain, France and Italy show the greatest inequality, and the West German structure is the most egalitarian; Belgium and the Netherlands lie between', and the relative pay of women varies greatly among these countries (Saunders and Marsden 1981, pp. 61, 238). Dispersion has also varied over time. Kuznets (1963) found a systematic relation between the extent of dispersion and the stage of development of the economy, dispersion increasing in the earlier stages of growth and then decreasing in the developed economy. Williamson and Lindert (1980) found that the course of change in the American pay structure agreed with this. The differential for skill was small before 1816, and then came a rapid widening, down to 1856. A further surge from 1899 brought the differential to its peak in 1916; but after 1929 a process of contraction set in and was maintained until the Korean war. The authors ascribed the change to three major factors. A high rate of investment displaced unskilled more than skilled labour, and was linked with a movement of labour out of agriculture. The uneven advance of productivity in different sectors affected the demand for skilled and unskilled labour differentially. Variations in immigration and fertility affected the relative supply of the unskilled. Elsewhere it has been pointed out that it is differences of pay between occupations that make up the greater part of dispersion, except in Great Britain, and attention has been directed to the impact on these differences of changes in demand and supply (Douty 1980, ch. 5, for USA). In particular, the extension of education has increased the relative supply of professional and technical qualifications. Trade union policy has taken effect, in Great Britain to reduce or eliminate the formerly very wide regional differences (Hunt 1973) and in Sweden, in pursuance of solidarity, to reduce the lead of builders' pay, raise the relative pay of women, and reduce the gap between white-collar and manual rates. Government policy has endeavoured to raise the lowest paid, as by the national minimum rate in France, Wages Councils in Great Britain, and the Fair Labor Standards Act in the USA, though the long-run effect on dispersion is uncertain. Incomes policies have affected differentials markedly in the short run. Other short-run changes arise from the trade cycle: rising activity has raised the lower rates relatively, and conversely.

Despite the variability of the pay structure in these ways, the regularity of its main features over

space and time remains outstanding. It challenges explanation, but remains a matter of discussion. Theories that have been advanced to account for the distribution of income are relevant here (Sahota 1978), as is the analysis of the log-normal form (Aitchison and Brown 1957). We may take it that pay is based on capacity. That the distribution of some measures of capacity such as IQ is normal and not log-normal is no bar to believing this, for it is understandable that as we go up the scale of capacity, what users are willing to pay rises more than proportionately, until we reach the vast earnings of those 'at the top of their profession'. We have then to explain why the distribution of capacity should take a common form in diverse societies. It seems likely that the explanation lies in the life-chances of the individual, and the impact of the myriad forces, beginning with conception, that shape body, mind, personality, training and experience. Though these forces have many different features in different societies, they share a stochastic property that gives a common form to the distribution of capacity, and hence to the inequality of pay.

See Also

- ▶ Discrimination
- ▶ Labour Economics
- ▶ Labour Markets
- ▶ Segmented Labour Markets

Bibliography

Aitchison, J., and J.A.C. Brown. 1957. *The lognormal distribution*. Cambridge: Cambridge University Press.

Bergson, A. 1984. Income inequality under Soviet socialism. *Journal of Economic Literature* 22(3): 1052–1099.

Douty, H.M. 1980. *The wage bargain and the labor market*. Baltimore/London: Johns Hopkins University Press.

Hunt, E.H. 1973. *Regional wage variations in Britain 1850–1914*. Oxford: Clarendon Press.

Kuznets, S. 1963. Quantitative aspects of the economic growth of nations. VIII: Distribution of income by size. *Economic Development & Cultural Change* 9(2): 1–79.

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

Mill, J.S. 1848. *Principles of political economy*. London: Parker.

Phelps Brown, E.H. 1977. *The inequality of pay*. Oxford: Oxford University Press.

Phelps Brown, E.H., and S.V. Hopkins. 1981. *A perspective of wages and prices*. London/New York: Methuen.

Rutter, M., and N. Madge. 1976. *Cycles of disadvantage*. London: Heinemann.

Sahota, G.S. 1978. Theories of personal income distribution: A survey. *Journal of Economic Literature* 16(1): 1–55.

Saunders, C., and D. Marsden. 1981. *Pay inequalities in the European community*. London: Butterworth.

Williamson, J.G., and P.H. Lindert. 1980. *American inequality, a macroeconomic history*. New York/London: Academic.

Infant Industry

Gerald M. Meier

Opposing arguments for free trade and protection constitute the longest-standing policy debate in the history of economic thought. In this debate the infant-industry argument has acquired pride of place as an exception to free trade – especially as trade theory now gives more attention to explicitly dynamic analysis instead of being confined to comparative statics. But the argument must be carefully stated, and when expressed in its precise modern form its applicability is narrowly limited.

During the period of mercantilism the argument was used to justify the granting of trade monopolies in new and hazardous trades and to inventions (Viner 1937, p. 71). Alexander Hamilton (1791), Friedrich List (1841) and J.S. Mill (1848) were also early prominent exponents of the argument. Since World War II it has acquired increasing emphasis for less developed countries.

The nature and scope of the infant-industry argument has been refined in modern times by the theory of domestic distortions and the application of welfare economics, with their concern for conditions of Pareto efficiency and determination of the cost of protection. It has also been delimited by considering the benefits and costs of alternative policy instruments – a subsidy, tariff

or quantitative restriction – in the context of the hierarchy of policy making (Bhagwati and Ramaswami 1963; Johnson 1965; Bhagwati 1971; Corden 1974).

The essence of the infant-industry argument rests on 'dynamic learning effects', so that the economy's transformation curve shifts outwards over time, and an industry that is not currently competitive may achieve comparative advantage after a temporary period of protection. Properly stated, the conditions necessary for infant-industry protection are: (1) irreversible technological external economies are generated that cannot be captured by the protected industry; (2) the protection is limited in time; and (3) the protection allows the industry to generate a sufficient decrease in economic costs such that the initial excess costs of the industry will be repaid with an economic rate of return equal to that earned on other investments.

If condition (1) is not fulfilled, the private market should be able to yield an efficient allocation unless capital markets are imperfect or there is imperfect information, so that risks are overestimated. Infant-industry protection is justified not by the fact that there are losses until the infant grows up – but by the fact of external economies associated with the learning process, so that there is underproduction from the social point of view. Condition (2) guarantees that the industry is not protected from infancy to geriatric or even senile stages. And condition (3) guarantees that the expected benefit must be sufficiently great to offset, in present value terms, the current costs of the policy required to produce the benefit (Kemp 1960).

If free trade is not optimal because of the presence of externalities and the possibility of lower costs over time, what then are the optimal policy instruments for protecting the infant industry? The normative theory of international trade policy has established that the first-best policy would be a production subsidy aimed at the source of the distortion (Corden 1974, pp. 28–31). This would be preferable to a tariff, which would lead to a by-product, consumption distortion. Although the tariff could restore equality between the marginal rate of domestic transformation and the marginal rate of transformation through foreign trade, it also would drive a wedge between the marginal rate of substitution in consumption and that of transformation. A tariff in turn would be preferable to a quantitative restriction, which would yield quota profits instead of customs revenue and would entail the cost of rent-seeking if there are import licences (Krueger 1974).

Although it is a domestic market failure that justifies the protection, nonetheless under certain types of market failure the first-best policy may not be a production subsidy (Corden 1984, pp. 91–2). If the learning experience results in dynamic internal economies in which the learning benefits remain wholly within the firm, the market failure may be in the imperfection of the capital market that makes the financing of such investment difficult or too expensive because the capital market is biased against this type of 'invisible' investment in human capital, or because the rate of interest for all long-term investment is too high owing to private myopia. In this case the first-best policy is to improve the capital market directly; a subsidy to that element of factor input or output that gives rise to the learning benefits would be second best, while further down the hierarchy there would be a general output subsidy to the industry, and then a tariff (Corden 1984, pp. 91–2).

Another case might involve dynamic external economies created by the labour training of a firm, but the firm is not able to retain the workers it has trained. In a perfect market situation the learning effects would be internalized: the workers would accept low wages during the learning stage, financing themselves by borrowing, with recoupment through subsequent mobility. But if the capital market is imperfect, or if there are rigidities in wage determination, this may not be possible. Again, the first-best policy is to improve the capital market; the second-best policy is to provide financing for, or subsidization to, the labour training; while subsidization of the firm's output would be further down the policy hierarchy.

Baldwin (1969) has indicated that a protective duty is no guarantee that individual entrepreneurs will undertake greater investments in acquiring technological knowledge. As long as the learning-

by-experience costs are higher than those which other firms must pay to acquire the knowledge, it cannot be assumed that firms will generally be prepared to incur the initial direct learning costs, even if the government imposes a tariff on the product. The duty will tend merely to encourage socially inefficient production as long as the state is willing to provide protection. A production subsidy on an industry-wide basis will have the same effect. What is needed is a direct and selective policy of subsidies to the initial entrants into the industry for discovering better productive techniques.

The infant-industry argument is also sometimes generalized to an 'infant economy' argument in which it is claimed that the entire industrial sector must go through an infancy stage, that the learning by each firm generates benefits for the whole sector and that by their mutual expansion all firms will enjoy a reduction in their production costs. Such a belief may underlie a broad import-substitution strategy with a uniform rate of effective protection to all manufacturing activities (Krueger 1984, p. 525). But import-substitution strategies beyond the first easy stage have proved excessively costly in developing countries, and their adverse effects on agricultural development and on export promotion have limited the rates of development in countries that have practised import-substitution protection (Balassa 1980).

In contrast to import substitution, it should be recognized that an export industry may also be an infant industry. Free trade may fail to bring about socially optimal levels of knowledge and factor endowment in new export industries. Policy interventions are then justified. Another possibility is that actual consumption experiences may be required to learn about an export commodity's qualities, but each firm's efforts at overcoming foreign-buyer resistance benefit not only itself but also all other firms that try to sell the same product in the same new market. The social returns of investments in market cultivation exceed the private returns, and subsidization is then justified (Mayer 1984). The higher rates of economic growth enjoyed by many countries that have promoted exports suggest that it is possible that the infant-industry proponents are correct in their basic argument that there is a period of learning and of relatively high costs, and that an export-promotion strategy is a more efficient way of developing an efficient, low-cost industrial structure (Krueger 1981, p. 16; Westphal 1981, p. 22).

Empirical evidence on infant-industry protection, however, is not as extensive as theoretical developments. Taussig (1888) concluded that there was legitimate application of protection 'for young industries' in the United States during the early period of 1789 to 1838. Marshall (1919), however, saw no clear evidence in support of intervention by the state in favour of nascent manufactures. For contemporary economies the empirical evidence with respect to infant industry protection is not definitive on its costs, benefits and duration of protection over time. Krueger and Tuncer (1982) showed that in Turkey there was no evidence to suggest that more protected industries experienced a higher rate of declining costs than less protected industries. The industries did not pass the necessary condition for an economic justification of protection, namely that they experienced more rapid gains in efficiency as judged by comparing domestic resource costs against foreign-exchange savings at shadow prices that properly reflect relative scarcities. Even though a protected industry may grow, the question remains whether it would not have grown in the absence of intervention. And the empirical question of potential benefits being greater than earlier costs must also be examined.

A major study concluded that productivity growth in infant industries appears to be highly variable and that few of the infant enterprises studied in less developed economies have demonstrated the high and continual productivity growth needed to achieve and maintain international competitiveness (Bell et al. 1984, p. 114). Moreover, high levels of protection have also tended to persist beyond a temporary learning period (ibid. p. 117), and there is little evidence that higher rates of protection have been given to industries with greater externalities.

Finally, regarding the degree of protection, Westphal (1981, p. 12) has suggested that – even for an 'efficient' infant industry, and evaluated at

prices that properly reflect relative scarcities – the domestic resource costs might initially be as much as twice the value of the foreign exchange saved or earned, with up to a decade being required to bring costs down to competitive levels. If production subsidies are given, the implied starting rate of subsidy in relation to value added is as much as 50 per cent. If, however, tariff protection is utilized, the rate of effective protection implied at the start of production is as high as 100 per cent.

Clearly the empirical justification for infant-industry protection will remain ambiguous until more research is done in quantifying the costs and benefits of protection, and its magnitude and duration. Such research is still in its own infancy.

See Also

- ▶ Comparative advantage
- ▶ Effective protection
- ▶ Industrialization
- ▶ National system
- ▶ Project evaluation
- ▶ Tariffs

Bibliography

Balassa, B. 1980. *The process of industrial development and alternative development strategies*, Essays in International Finance No. 141. Princeton: Princeton University, December.

Baldwin, R.E. 1969. The case against infant-industry tariff protection. *Journal of Political Economy* 77(3): 295–305.

Bardhan, P.K. 1970. *Economic growth, development and foreign trade*. New York: Wiley.

Bell, M., B. Ross-Larson, and L.E. Westphal. 1984. Assessing the performance of infant industries. *Journal of Development Economics* 16: 101–128.

Bhagwati, J. N. 1971. The generalized theory of distortions and welfare. In *Trade, balance of payments and growth*, ed. J.N. Bhagwati, et al. Amsterdam: North-Holland, ch. 12.

Bhagwati, J.N. 1978. *Foreign trade regimes and economic development: Anatomy and consequences of exchange control regimes*. New York: National Bureau of Economic Research.

Bhagwati, J.N., and V.K. Ramaswami. 1963. Domestic distortions, tariffs and the theory of optimum subsidy. *Journal of Political Economy* 71: 44–50.

Clemhout, S., and H.Y. Wan. 1970. Learning-by-doing and infant industry protection. *Review of Economic Studies* 37: 33–56.

Corden, W.M. 1971. *The theory of protection*. Oxford: Clarendon Press.

Corden, W.M. 1974. *Trade policy and economic welfare*. Oxford: Clarendon Press.

Corden, W.M. 1984. Normative theory of international trade. In *Handbook of international economics*, vol. 1, ed. R.N. Jones and P.B. Kenen. Amsterdam: North-Holland.

Hamilton, A. 1791. *Report on manufactures*. Reprinted in US Senate Documents XXII/172, Washington, DC: Congress, 1913.

Johnson, H.G. 1965. Optimal trade intervention in the presence of domestic distortions. In *Trade, growth, and the balance of payments*, ed. R. Caves, H.G. Johnson, and P.B. Kenen. New York: Rand McNally.

Johnson, H.G. 1970. A new view of the infant industry argument. In *Studies in international economics: Monash conference papers*, ed. A. McDougall and R.H. Snape. Amsterdam: North-Holland.

Kemp, M.C. 1960. The Mill–Bastable infant-industry dogma. *Journal of Political Economy* 68(February): 65–67.

Krueger, A.O. 1974. The political economy of the rent-seeking society. *American Economic Review* 64: 291–303.

Krueger, A.O. 1978. *Foreign trade regimes and economic development: Liberalization attempts and consequences*. Cambridge, MA: Ballinger for the National Bureau of Economic Research.

Krueger, A.O. 1981. Export led industrial growth. In *Trade and growth of the advanced developing countries in the Pacific Basin*, ed. W. Hong and L.B. Krause. Seoul: Korea Development Institute.

Krueger, A.O. 1984. Trade policies in developing countries. In *Handbook of international economics*, vol. I, ed. R.W. Jones and P.B. Kenen. Amsterdam: North-Holland.

Krueger, A.O., and B. Tuncer. 1982. An empirical test of the infant industry argument. *American Economic Review* 72(5): 1142–1152.

List, F. 1841. *Das nationale System der Politischen Oekonomie.* Jena: Gustav Fischer, 1920. Trans. by G.P.A. Matile as *National system of political economy.* Philadelphia: Lippincott, 1856.

Little, I.M.D., T. Scitovsky, and M.F.G. Scott. 1970. *Industry and trade in some developing countries: A comparative study.* London: Oxford University Press.

Marshall, A. 1919. *Industry and trade.* London: Macmillan, Appendix G, 2.

Mayer, W. 1984. The infant-export industry argument. *Canadian Journal of Economics* 17(May): 249–269.

Meade, J.E. 1955. *Trade and welfare*. New York: Oxford University Press.

Mill, J.S. 1848. *Principles of political economy*, ed. W.-J. Ashley. London: Longmans Green, 1909.

Taussig, F.W. 1888. *The tariff history of the United States*. New York: G.P. Putnam's Sons.

Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper & Bros.
Westphal, L.E. 1981. *Empirical justification for infant industry protection*, World Bank Staff Working Paper No.445. Washington, DC: World Bank, March.

Infant Mortality

K. Wolpin

There is extensive variation in the level of infant mortality (deaths under 1 year of age) across countries, over time within countries, and across subgroups within countries or regions. Social scientists, and demographers in particular, have devoted a great deal of research effort towards identifying the underlying sources of variation, biological and genetic, environmental and behavioural, and their relative importance.

To gain perspective on the degree of cross-sectional and temporal variation in infant mortality rates, the following data are useful. According to World Fertility Survey statistics obtained in the mid-1970s, in Bangladesh 13.5% of all children ever born died before the age of one, in Mexico 7.2% of infants died, and in Malaysia 3.6%. Comparable figures were approximately 1.9% for the USA, 1.3% for Japan, and 1.2% for Sweden. Around 1900 the USA had an infant mortality rate equal to that cited for present-day Bangladesh; around 1925 it was equal to that of present-day Mexico; and directly after World War II to that in present-day Malaysia. Further, within the region of West Africa, a high mortality area, the infant mortality rate in 1972 varied from 12.2% in Ghana to 21.6% in Guinea.

It is not my intention to review the evidence on the relative importance of the various factors thought to influence the level of infant mortality, for in my view that literature has serious methodological flaws stemming from the lack of a consistent theoretical paradigm. Suffice it to say that there is still much debate even among those who would not hold this view. I wish instead to present an economic perspective which has a critical bearing on the methodological approach used to resolve this debate. This perspective draws heavily on the notion of household production (Becker 1965).

The essence of this approach is that infant health, survival being one albeit very important aspect of health, is produced according to some technological function which includes as inputs the resources devoted to the child (during pregnancy and after birth), such as prenatal care, breastfeeding, parental time devoted to child care, vaccinations, etc.; environmental conditions, such as sanitation and weather; biological conditions, such as the interval between births, the age of the mother at birth, and the genetic endowment of the child. More precisely, these inputs yield a probability distribution over health outcomes. Infant mortality differs systematically among individuals within a society, across societies, and over time within societies because of the differences in the levels and mix of these inputs and because of differences in the characteristics of the technology. In order to estimate the production function from data of any kind, it is necessary to postulate a mechanism which accounts for the input variation in the data.

From the economist's perspective it is natural to think of individuals as optimizing subject to constraints. Inputs have prices, monetary and/or psychic; faced with a given health technology, input levels are chosen depending upon the array of input prices both currently and the distribution expected to prevail in the future, wage rates (current and future), other income sources (current and future), preferences, and family and child endowments not subject to choice to the extent that they are known by the household. Much of the empirical literature can be seen as attempting to estimate technology, although in most of that literature technology is confounded with preferences through the introduction of income or prices. Further, few studies have estimated technology, accounting for the fact that, within an optimization framework, input choices would be conditioned on family or child endowments and on exogenous environmental factors. For example, because an infant's intake of breast milk depends on its ability to suckle, immature or ill infants may thus be breastfed less or not at all,

leading to an upward bias in the estimation of the effect of breastfeeding on infant health or survival. A number of recent papers which adopt this household production framework using both micro-level data from developed (Rosenzweig and Schultz 1983) and less developed (Olsen and Wolpin 1983) countries have demonstrated the importance of the assumptions about the process generating input variation in estimating input effects.

It is plausible that infant mortality is linked behaviourally to other demographic decisions, such as the number, timing, and spacing of children. Indeed, there is a large literature which has posed the question about the impact of infant deaths on fertility, presuming the variation in infant mortality at the individual level to arise solely or mostly from stochastic events not subject to control. Two distinct fertility strategies have been discussed – replacement and hoarding. Replacement refers to the fertility reaction to a realized death, while hoarding refers to a strategy of acquiring an inventory of children in anticipation of future deaths (Ben-Porath 1976; Schultz 1976). Replacement behaviour would arise in the simplest of dynamic models with infant survival uncertainty because an infant death must increase the marginal benefits of an additional child. Hoarding behaviour, which is a response to the *ex ante* survival uncertainty, will arise only if mortality of older children is significant and/or surviving children are desired early in the life cycle. Although it is recognized that this behaviour, if rigorously modelled, would require solving a complicated dynamic stochastic optimization problem, most attempts to estimate these effects have been statistical in nature (Olsen 1980) and only loosely based on theory. Some estimates of replacement based explicitly on a behavioural formulation have been obtained (Wolpin 1984), but solving and estimating such models is computationally burdensome. Formulating and estimating a dynamic model with hoarding is a more ambitious undertaking than has yet been accomplished. Incorporating health investment decisions in children in a dynamic choice setting has yet to be implemented, although that is where this literature is and ought to be moving.

But, what does the household choice theoretical framework have to do with the enormous differentials in infant mortality we observe between countries and the historically extraordinary decline in infant mortality throughout the world? Surely the individuals in Bangladesh cannot choose to have an infant mortality rate equal to that in the USA. One can view this question in two ways. At a superficial level, it is clear from the figures cited above that infant mortality is inversely related to per-capita income in the cross-section and the time series. For example, using data from the World Fertility Survey countries, the per-capita income for countries with an infant mortality rate above 10% was around $350, for countries with an infant mortality of between 7.5 and 10%, it was $600, for countries with an infant mortality rate between 5.0 and 7.5%, $1302, and for countries with infant mortality rates between 2.5 and 5.0%, $2168. Of course, it is not income per se which causes reductions in infant mortality, but the improved preventive medicine and eradication of disease, the introduction of modern sanitation, and the improved food distribution, which come with economic development. Even ignoring the fact that the relationship between infant mortality and income is far from perfect – for example, Turkey had in the 1970s an infant mortality rate similar to that in Bangladesh but a per-capita income level ten times as large – the relationship between infant mortality and income is not fundamental. To the extent that innovations in medicine and the like parallel changes in economic circumstances of the population as a whole and are intertwined with the fundamental desires of the population, they too require explanation in the context of overall economic and social development. What one learns from the household choice framework is that to understand the cross-section and time-series aggregate data, what is needed is a model of economic growth which incorporates endogenous demographics, by which I mean conscious choice about investments in human capital (infant health, for example) and family size, in addition to investments in physical capital. Introducing modern sanitation, for example, requires real resources in research and implementation and some decision process must be responsible for the diversion of resources to that use.

To Malthus (1798), mortality played a crucial role as a positive check on population growth. One can always argue about what Malthus really meant, but stripped to essentials the argument was that the fixed capacity of land coupled with exogenous population growth causes consumption per capita to fall to the subsistence level. Although Malthus recognized that fertility might respond to falling living standards, this preventive check was in his view weak. The equilibrium mortality rate is that rate which constrains population size to remain at the level consistent with steady state subsistence consumption.

The standard neoclassical growth model, e.g. Solow (1956), takes net population growth (fertility net of mortality) as exogenous. Per-capita consumption is maximized when the marginal product of capital is equal to the net reproduction rate. This model leads to the result that the 'optimal' mortality rate is equal to the excess of the fertility rate over the replacement fertility rate. Samuelson (1975) noticed that in an overlapping generations growth model, the optimal rate of growth of population is infinite because welfare rises continously the more young there are to support the old. Samuelson conjectured that there would be a deterministic optimum for population in a model which combined the neoclassical and overlapping generations approaches. Unfortunately, an overlapping generations model with capital (Diamond 1965) does not yield an interior solution for the optimal population growth rate for any unbounded production function (Deardorff 1976); again, population growth should optimally be zero. Because the overlapping generations framework is based on individual optimization, it allows for a broadening of the choice set. It does not make a great deal of sense to discuss optimal population issues in models where there exists no mechanism to achieve the optimum. There have been some attempts to allow for endogenous population in such models by adopting the notion from the microeconomic literature on fertility that households can choose their fertility, where children are consumption goods requiring expenditures (Eckstein and Wolpin 1985; Nerlove et al. 1984). These models are capable of describing the time path of output, consumption and population, and can yield insights as to the impact of technology, taste and endowments on those time paths. Incorporating infant mortality into overlapping generations growth models as a choice outcome, in the sense of allowing for investment in the human and physical capital necessary to affect it, would seem to be a logical and important step forward in understanding the economic and demographic development process. In particular, in many now developed countries, infant mortality decline preceded the decline in fertility so that population growth increased during the transition. Can an economic growth model in which fertility and investments in child survival are explicitly chosen by economic agents account for such a demographic pattern, and can this be fit into the observed timing of economic growth?

The notion that infant mortality, more broadly infant health and even more broadly child human capital, is and has always been an economic decision in a fundamental sense forces one to think more carefully about the determinants of aggregate cross-sectional and time-series demographic and economic variables. Much of this work is itself in an infant stage; its health and survival is also subject to choice.

See Also

- Demographic Transition
- Fecundity
- Fertility
- Nutrition
- Mortality
- Public Health

Bibliography

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75, September: 493–517.

Ben-Porath, Y. 1976. Fertility response to child mortality: Micro data from Israel. *Journal of Political Economy* 84, August: S163–S178.

Deardorff, A. 1976. The optimum growth rate for population: Comment. *International Economic Review* 17(2), June: 510–515.

Diamond, P. 1965. National debt in a neoclassical growth model. *American Economic Review* 55, December: 1126–1150.

Eckstein, Z., and K. Wolpin. 1985. Endogenous fertility and optimum population size. *Journal of Public Economics* 27, June: 93–106.
Malthus, T. 1798. *An essay on the principle of population.* ed. A. Flew, Harmondsworth: Penguin, 1970.
Nerlove, M., Razin, A., and E. Sadka. 1984. Income distribution policies with endogenous fertility. *Journal of Public Economics* 24, March: 221–230.
Olsen, R. 1980. Estimating the effects of child mortality on the number of births. *Demography* 17, November: 429–443.
Olsen, R., and K. Wolpin. 1983. The impact of exogenous child mortality on fertility: a waiting time regression with dynamic regressors. *Econometrica* 51, May: 731–749.
Rosenzweig, M., and T.P. Schultz. 1983. Demand for health inputs, and their effects on birth weight. *Journal of Political Economy* 91, October: 723–746.
Samuelson, P. 1975. The optimum growth rate of populations. *International Economic Review* 16, October: 531–538.
Schultz, T.P. 1976. Interrelationships between mortality and fertility. In *Population and development: The search for selective interventions*, ed. R. Ridker. Baltimore: Johns Hopkins Press.
Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, February: 65–94.
Wolpin, K. 1984. An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy* 92, October: 852–874.

Infant-Industry Protection

Douglas A. Irwin

Abstract

The infant industry argument for protection holds that new industries in developing countries should be promoted through trade or industrial policy measures to allow them to mature and compete successfully against established foreign rivals. This article assesses the theory and evidence behind this qualification to the case for free trade.

Keywords

Brazil; Capital market failure; Cost–benefit analysis; Counterfactual simulation; Denmark; Dynamic learning effects; Growth and learning-by-doing; Hamilton, A.; Imperfect information; Infant industry protection; Informational barriers to entry; Learning-by-doing; Learning-by-doing spillovers; List, F.; Mill, J. S.; Research and development; Smith, A.; Trade policy, political economy of; Transfer of technology; Underproduction

JEL Classifications

F1

The infant industry argument for trade protection is one of the oldest and most widely debated qualifications to the case for free trade. The argument holds that certain new industries should be protected from foreign competition in the expectation that they will eventually mature and successfully compete against more experienced foreign rivals. The case for infant industry protection involves temporary and selective, not permanent and across-the-board, government assistance and is often discussed in the context of trade policies that might promote economic development.

The idea of infant industries can be traced back as far as the 17th century (Irwin 1996). In Book IV of the *Wealth of Nations* (1776), Adam Smith was sceptical that trade restrictions would create new wealth, arguing that they would just divert scarce resources into less productive endeavours. Other writers, such as Alexander Hamilton (1791) and Friedrich List (1841), believed that policies to promote manufacturing industries would be beneficial in encouraging economic diversification and growth in developing countries. The classical economist John Stuart Mill (1848, p. 922) lent his authority to the case by endorsing it in this way:

> The only case in which, on mere principles of political economy, protecting duties can be defensible, is when they are imposed temporarily (especially in a young and rising nation) in hopes of naturalizing a foreign industry, in itself perfectly suitable to the circumstances of the country. The superiority of one country over another in a branch of production, often arises only from having begun it sooner. There may be no inherent advantage on one part, or disadvantage on the other, but only a present superiority of acquired skill and experience. A country which has this skill and experience yet to acquire, may in other respects be better adapted to

> the production than those which were earlier in the field.... But it cannot be expected that individuals should, at their own risk, or rather to their certain loss, introduce a new manufacture, and bear the burthen of carrying it on until the producers have been educated up to the level of those with whom the processes are traditional. A protecting duty, continued for a reasonable time, will [*changed to* '*might*' *in later editions*] sometimes be the least inconvenient mode in which the nation can tax itself for the support of such an experiment. But ['*it is essential that*' *added in later editions*] the protection should be confined to cases in which there is good ground of assurance that the industry which it fosters will after a time be able to dispense with it; nor should the domestic producers ever be allowed to expect that it will be continued to them beyond the time necessary for a fair trial of what they are capable of accomplishing.

In the 19th century, the debate over infant industry protection centred on whether such protection would (*a*) create new wealth and capital, or merely divert it from other more profitable activities, (*b*) stimulate domestic producers to acquire new technology and skills, or just stifle the incentive for such efforts, and (*c*) generate long-term net benefits, or simply foster costly industries that would require ongoing government support. Unfortunately, economic analysis proved to be of little assistance in evaluating these claims, as one could envision the successful maturation of an infant industry but also see the possibility of protection breeding inefficiencies; a priori, neither outcome could be dismissed.

In the modern literature, the infant industry argument hinges on dynamic learning effects, which allow an industry that is not currently competitive to become so after a temporary period of protection. As such, the conditions for infant industry protection include the following: (*a*) irreversible technological external economies that cannot be captured by the protected industry, (*b*) a limited period of protection, and (*c*) sufficient long-run economic benefits (lower production costs that generate producer surplus) that will more than compensate for the costs associated with protection, with a rate of return at least equal to that on other investments (Kemp 1960).

If condition (*a*) is not fulfilled, the private market should deliver an efficient outcome unless there is some other market imperfection (poorly functioning capital markets, imperfect information) so that risks to starting the industry are overestimated by private agents and there is underproduction from the social point of view. Condition (*b*) states that protection must be time-limited, and not persist indefinitely. Condition (*c*) requires an intertemporal cost–benefit analysis, wherein the initial costs of protecting the industry will be more than offset by long-run benefits.

The modern literature on infant industries also focuses on identifying the specific market failure or distortion that makes government intervention necessary as well as the ranking of alternative policy instruments in terms of their ability to correct the market failure or distortion. If an industry is characterized by learning-by-doing spillovers, wherein production costs for any firm fall as a result of production experience by anyone in the industry (that is, the learning benefits are external to the firm), then this dynamic economy of scale may lead to a divergence between private and social costs of production. The knowledge generated by research and development expenditures can also create external benefits which could lead to underinvestment by the private market. And if there are capital market failures, such that firms cannot acquire credit (Flam and Staiger 1991; Bond 1993), or informational barriers to entry (Grossman and Horn 1988), there may be a case for selective interventions.

Once the specific obstacle facing the infant industry is identified (that is, related to production experience, technology transfer, or imperfect capital markets), then the policy recommendation can be specifically targeted to address the problem. In general, trade protection will not be the first-best policy intervention to correct the distortion that hinders the development of an infant industry. Baldwin's (1969) classic critique of the infant industry argument stresses that import protection fails to provide the right incentive for an infant firm to make additional investments in acquiring technological knowledge and does not necessarily solve a firm's appropriability problem of securing the benefits of investments in knowledge or production experience. But, by reducing foreign competition and raising the domestic price, it does make the status quo more profitable.

If the economic conditions giving rise to infant industries are difficult to assess, the implementation of a welfare-improving policy also poses difficulties for governments. The government must differentiate among various industries (ignoring the lobbying of firms for government assistance), pick those to support with preferential policies, select the proper policies to ensure that firms have the incentive to respond the right way, and be able to resist pressure from firms to maintain protection indefinitely (Tornell 1991).

Despite many theoretical articles on infant industry protection, Krueger and Tuncer (1982, p. 1142) wrote that 'there has been virtually no systematic examination of the empirical relevance of the infant industry argument' through *ex post* evaluations or other studies. They found little correlation between various measures of the effective rate of protection and industry productivity growth in Turkey, but did not perform counterfactual simulation or cost–benefit analysis (see Harrison 1994).

To date, there are still relatively few evaluations of infant industry policies. Luzio and Greenstein (1995) studied performance of the Brazilian microcomputer industry under protection that started in the early 1990s. They found that the rates of technological advance in Brazil were rapid but lower than that of potential international competition. As a result, the technical frontier in Brazil lagged that best performance practices in international markets by three to five years, and forgone consumer surplus due to protection approached 20 per cent of domestic expenditure on microcomputers.

Hansen et al. (2003) examined the welfare effects of subsidies in Denmark for the production of electricity from wind power. They found strong learning-by-doing productivity growth in the Danish windmill industry, and the industry achieved a dominant position in the world market. By their calculation, these subsidies passed a cost–benefit test: the costs consist of the efficiency loss from diverting electricity production from using fossil fuels to utilizing wind power, but the benefits include reductions in the environmental damage of using fossil fuels and the emergence of a new export sector. They concluded that the subsidies pass a cost–benefit test because the value of the windmill firms at the stock exchange far exceeds the accumulated distorted losses in electricity production.

Economists have also looked at historical cases of infant industry protection, as it is commonly contended that high-income countries such as the United States, Japan and Germany rose to industrial prominence by protecting infant industries in the late 19th and early 20th centuries (Chang 2003). Two studies examined the US iron and steel industry in the late 19th century. Head (1994) sought to account for the individual roles of learning-by-doing, changing resource endowments, and tariff protection in the emergence of the steel rail industry. In a counterfactual simulation of what would have happened under free trade, he concluded that learning effects were very strong and that, even though the steel rail tariff hurt rail users in both short and long runs, the tariff's overall effect on welfare was positive but fairly small. Irwin (1998) studied the US tinplate industry which, after earlier failures, flourished after receiving tariff protection in 1890. His counterfactual simulation indicated that, without the additional duties, domestic tinplate production would have arisen about a decade later as US iron and steel input prices converged to those in Britain. Although the tariff accelerated the industry's development, welfare calculations suggest that protection did not pass a cost–benefit test.

In general, however, economists have been sceptical about the relevance of the infant industry argument for current developing countries, and for the ability of governments to implement the policy wisely. For example, reviewing the empirical literature on manufacturing establishments in developing countries, Tybout (2000) found that unexploited economies of scale in developing countries are insignificant and that protection tends to reduces average efficiency levels by allowing lower-productivity, higher-cost firms to survive in the market. He concluded that 'although the econometric evidence on technology diffusion in [developing countries] is limited, it does suggest that protecting "learning" industries is unlikely to foster productivity growth' (2000, p. 39).

See Also

- ▶ Growth and International Trade
- ▶ International Trade Theory
- ▶ Strategic Trade Policy
- ▶ Trade Policy, Political Economy of
- ▶ Trade, Technology Diffusion and Growth

Bibliography

Baldwin, R. 1969. The case against infant-industry tariff protection. *Journal of Political Economy* 77: 295–305.

Bell, M., B. Ross-Larson, and L. Westphal. 1984. Assessing the performance of infant industries. *Journal of Development Economics* 16: 101–128.

Bond, E. 1993. Capital market imperfections and the infant industry argument for protection. In *Theory, policy and dynamics in international trade: Essays in honor of Ronald W. Jones*, ed. W. Ethier, E. Helpman, and J. Neary. New York: Cambridge University Press.

Chang, H.-J. 2003. Kicking away the ladder: Infant industry promotion in historical perspective. *Oxford Development Studies* 31: 21–32.

Flam, H., and R. Staiger. 1991. Adverse selection in credit markets and infant industry protection. In *International trade and trade policy*, ed. E. Helpman and A. Razin. Cambridge, MA: MIT Press.

Grossman, G., and H. Horn. 1988. Infant-industry protection reconsidered: The case of informational barriers to entry. *Quarterly Journal of Economics* 103: 767–788.

Hamilton, A. 1791. Report on manufactures. In *The papers of Alexander Hamilton*, ed. H. Syrett. New York: Columbia University Press 1961.

Hansen, J., D. Jensen, and E. Madsen. 2003. The establishment of the Danish windmill industry – was it worthwhile? *Review of World Economics* 139: 324–347.

Harrison, A. 1994. An empirical test of the infant industry argument: Comment. *American Economic Review* 84: 1090–1095.

Head, K. 1994. Infant industry protection in the steel rail industry. *Journal of International Economics* 37: 141–165.

Irwin, D. 1996. *Against the tide: An intellectual history of free trade*. Princeton: Princeton University Press.

Irwin, D. 1998. Did late-nineteenth-century US tariffs promote infant industries? Evidence from the tinplate industry. *Journal of Economic History* 60: 335–360.

Kemp, M. 1960. The Mill–Bastable infant-industry dogma. *Journal of Political Economy* 68: 65–67.

Krueger, A., and B. Tuncer. 1982. An empirical test of the infant industry argument. *American Economic Review* 72: 1142–1152.

List, F. 1841. *National system of political economy*. Philadelphia: J. Lippincott.

Luzio, E., and S. Greenstein. 1995. Measuring the performance of a protected infant industry: The case of Brazilian microcomputers. *Review of Economics and Statistics* 77: 622–633.

Mill, J.S. 1848. *Principles of political economy*, 1909. London: Longmans.

Smith, A. 1776. *The wealth of nations*, 1976. Oxford: Clarendon Press.

Tornell, A. 1991. Time inconsistency of protectionist programs. *Quarterly Journal of Economics* 106: 963–974.

Tybout, J. 2000. Manufacturing firms in developing countries: How well do they do, and why? *Journal of Economic Literature* 38: 11–44.

Inflation

Michael Parkin

Abstract

This article essay reviews the theoretical and empirical literature on the causes and consequences of inflation – of a continuously rising price level and falling value of money. It describes the research agendas using the analytical distinction between anticipated inflation – an idealized situation in which prices are rising at a rate at which all economic agents expect them to rise – and unanticipated inflation. The literature on the effects of inflation on economic growth and unemployment, inflation in open economies, positive theories of central bank behavior, inflation and fiscal policy, and policies towards inflation including interest rate and inflation targeting receives particular attention.

Keywords

Aggregate demand shocks; Budget deficits; Business cycles; Capital–labour ratio; Cash-in-advance constraint; Central bank behaviour; Commodity money; Consumer Price Index; Dynamic general equilibrium analysis; Exchange rate determination; Exchange rate regimes; Fiscal theory of the price level; Hyperinflation; Incomplete information; Inflation; Inflation measurement; Inflation targeting; Interest rate rules; International policy coordination; Labour market contracts;

Labour unions; Monetarism; Monetary and fiscal policy; Monetary base; Monetary policy rules; Money; Money supply; Monopolistic competition; Mundell–Tobin effect; Natural rate of unemployment; Neutrality of money; New classical theory; New Keynesian macroeconomics; Output gap; Overlapping generations framework; Overnight interest rate on inter-bank loans; Phillips curve; Precommitment; Prices and incomes policies; Rational expectations; Real business cycles; Reputation; Sacrifice ratios; Stagflation; Sticky prices; Sticky wages; Stylized facts; Substitutes and complements; Superneutrality of money; Targets and instruments; Technological shocks; Time inconsistency; Time preference; Velocity of circulation

JEL Classifications
E3

'Inflation is a process of continuously rising prices, or equivalently, of a continuously falling value of money' (Laidler and Parkin 1975, p. 741). Because there are several ways of measuring prices, there are also several different measures of inflation. The most commonly used measures in the modern world are the percentage rate of change in a country's Consumer Price Index or in its Gross Domestic Product deflator. Measures of inflation in earlier periods are based on fragmentary samples of prices, such as those of corn and other staple commodities, or of labour.

Inflation has been a feature of human history for as long as money has been used as a means of payment, and as Milton Friedman (1970, p. 24) famously wrote, 'inflation is always and everywhere a monetary phenomenon, in the sense that it cannot occur without a more rapid increase in the quantity of money than in output'.

Anna J. Schwartz (1973) provides a compact account of the history of inflation from antiquity to modern times. One of the earliest documented inflations in the ancient world occurred following Alexander the Great's conquest of the Persian Kingdom (330 BC); the Roman Empire experienced rapid inflation under Diocletian at the end of the third century AD. We have no knowledge of inflation for the thousand years that followed the fall of the Roman Empire. But we do have data from the Middle Ages onwards. The inflation episodes during the Middle Ages were modest, and during those years there was a tendency for periods of rising prices to be interspersed by periods of falling prices. This pattern of intermittent inflation and deflation persisted all the way through to the Great Depression of the 1930s. Since the Great Depression, there has been a general tendency for prices to rise every year (with trivial exceptions). In the 1970s and early 1980s, serious inflations – of more than ten per cent a year – gripped most of the industrial world. But this 'double-digit' inflation era was short-lived, and by the mid-1980s inflation rates had returned to the more modest levels experienced in the late 1960s. In the early 2000s, there was little sign of high inflation returning in the major economies. Individual inflations of spectacular dimensions occurred in inter-war Europe, during the fall of Nationalist China (1948–9), and in modern times in some Latin American nations, Israel, and Zimbabwe. Some of these were episodes were hyperinflations – inflation rates that exceeded 50 per cent per month.

It is the fact that inflation has been so variable over time and across countries that gives rise to the question: what are the causes and the consequences of inflation? It is the enormously rich variation in inflationary experience that also provides the data which makes progress in answering those questions possible.

The literature on inflation is large, and several comprehensive, if dated, surveys of it are available (see Bronfenbrenner and Holzman 1963; Johnson 1963; Laidler and Parkin 1975). No up-to-date survey of the literature on inflation was available as of 2006.

Attempts to understand inflation have been aided by the insight that anticipated inflation has different effects from unanticipated inflation. It is convenient to use that distinction in organizing this article. But it must be borne in mind that the distinction between anticipated and unanticipated inflation is analytical. It is not a distinction that

has an immediate or direct correspondence with actual historical inflations.

Anticipated Inflation

Anticipated inflation is an idealized situation in which prices are rising at a rate at which all economic agents expect them to rise. No one is caught by surprise. What are the effects of a fully anticipated inflation?

There is little disagreement on the answer to this question concerning the effects on nominal variables – on such things as nominal interest rates, wages and foreign-exchange rates. Other things equal, the higher the expected rate of inflation, the higher the *level* of nominal interest rates, the higher is the rate at which wages rise, and the faster the rate of currency depreciation. Furthermore, these effects are one for one. An x per cent higher anticipated inflation raises nominal interest rates by x per cent, makes wage rates rise x per cent faster, and makes the currency depreciate x per cent faster.

There is less than complete agreement about the effects of anticipated inflation on *real* economic variables. Abstracting from transitory adjustment paths, all economic theories predict monetary neutrality: a one-shot change in the quantity of money leads to a proportionate change in the levels of all prices (and wages) and has no real effects. But not all economic theories predict monetary superneutrality – that real variables are neutral with respect to changes in the growth rate of the quantity of money.

There are three alternative views in the literature concerning money's superneutrality. One view is that money is superneutral – a change in the anticipated inflation rate has no effects on output (or economic welfare). A second view is that in increase in the anticipated inflation rate increases output (and economic welfare). Yet a third view is that a higher anticipated inflation rate lowers output (and economic welfare).

The superneutrality result has been most elegantly and clearly stated by Sidrauski (1967). The result also is present in some modern theories of money that pay detailed attention to the physical environment in which monetary exchange arises (see, for example, Townsend 1980). The essential feature of models that generate superneutrality is that the real rate of interest is imposed by the structure of preferences (intertemporally additive with a constant rate of time preference). In equilibrium, the marginal product of capital is equal to this fixed rate of time preference so that, regardless of what happens to money, the capital stock and output rate are unaffected.

The natural rate hypothesis is a variant of the superneutrality proposition. This hypothesis, advanced by Friedman (1968) and Phelps (1968), states that money is superneutral in the particular sense that there is a unique natural unemployment rate that is independent of the anticipated rate of inflation. Any trade-off between inflation and unemployment is temporary and best thought of as a trade-off between unanticipated inflation and unemployment.

The second view that a higher anticipated rate of inflation increases output and improves economic welfare arises in two classes of models. The first is the so-called Mundell–Tobin effect (Mundell 1963, 1965; Tobin 1965). A higher anticipated inflation rate results in an increase in the opportunity cost of holding real money balances. According to the Mundell–Tobin view, this higher opportunity cost of holding money leads to a portfolio reallocation away from money and towards physical capital. The higher holdings of physical capital result in a higher stock of capital and therefore in a higher capital–labour ratio, which in turn leads to a higher level of output. A rise in the anticipated rate of inflation would put the economy on an adjustment path towards the new higher capital stock that would be associated with a transitory rise in the growth rate and a permanent rise in the level of output. A restatement of the Mundell–Tobin position couched in a modern rational expectations terms has been provided by Fischer (1979). The second type of model is one in which an asymmetry in price and wage adjustment – a downward rigidity – creates a long-run trade-off between inflation and the level of economic activity – a downward-sloping long-run Phillips curve.

The third view that a higher anticipated rate of inflation lowers output and economic welfare also arises in two classes of models. First in an overlapping-generations framework (Samuelson 1958; Wallace 1980) a rise in the anticipated rate of inflation leads agents to economize on their holdings of money which, in turn, leads them to save less and transact on a lower scale with the succeeding generation. Second, Clower's (1967) suggested technological basis for money – the cash-in-advance constraint – generates super-non-neutrality. Using Clower's assumption, Stockman (1981) shows that, because a higher anticipated inflation rate raises the opportunity cost of holding money, this, in effect, raises the opportunity cost of undertaking all transactions and, therefore, in equilibrium lowers the scale of transactions undertaken. In Stockman's model, this results in a lower investment rate and lower capital stock. Thus a higher expected inflation rate leads to a lower level of output. A rise in the anticipated inflation rate will place the economy on an adjustment path that would result in a lower transitory growth rate and a lower permanent level of income.

Some of the above results can be thought of in terms of the substitute/complement relation between money and capital. If money and capital are substitutes in portfolios, then the Mundell–Tobin result arises. If money and capital are complements, as they implicitly are in the overlapping generations and cash-in-advance models, then higher anticipated inflation leads to lower output.

There is an abundance of empirical evidence on the alternative hypotheses about the effects of fully anticipated inflation. But the evidence is not entirely unambiguous. Because the very concept of anticipated inflation is analytical and not historical, in examining inflationary experience assumptions must be made concerning the extent to which inflations have been anticipated.

Comprehensive and systematic attempts that have addressed the question in the context of economic growth are those by Kormendi and Meguire (1985), Barro (1997), and Sala-i-Martin et al. (2004).

Using post-war data for 47 countries, Kormendi and Meguire analyse the effects of a change in the anticipated rate of inflation on output growth in a multivariate regression framework. Anticipated inflation was measured as simply the mean growth rate of inflation over the sample period (which went from the late 1940s to 1977). The finding of that study solidly rejects the Tobin–Mundell hypothesis and, in some formulations, fails to reject the opposite view.

Using data for about 100 countries between 1960 and 1990, Robert Barro finds that inflation has a negative effect on growth. The effect is small but significant and implies that maintained for a number of years, in inflation rate that exceeds ten per cent per year has a large cumulative effect on output. Barro is careful in his analysis of the endogeneity of inflation and growth to establish that causation runs from inflation to growth.

Barro's finding is challenged by Sala-i-Martin, Doppelhofer and Miller. Using data from 1960 to 1996 for 88 countries and 67 variables considered candidates for influencing the rate of economic growth, and using a Bayesian averaging of classical estimates approach, they find that neither average inflation rate nor the square of the inflation rate has a significant effect on the growth rate.

The work just summarized takes a reduced form and linear approach, and these features limit its utility. Future work on the effects of inflation on growth should be directed toward looking at structural accounts of the linkages and seeking highly nonlinear and perhaps nonparametric relationships between these two variables.

Investigations of the neutrality of unemployment (and output) with respect to anticipated inflation has been the subject of innumerable studies, and Laidler and Parkin (1975) review the state of this literature up to the mid-1970s. The conclusions that emerged from this work were mixed, and most of the results generated on data-sets that ended around 1970 showed the existence of a trade-off. But as the data for the 1970s (with its high inflation rate) were added, the picture changed and Laidler and Parkin concluded that it was not possible to reject the view that the unemployment rate is neutral with respect to anticipated inflation.

This conclusion is challenged in three different ways. First, the classic Sargent (1976) shows that reduced-form equations estimated for a given sampling interval over a given sampling period cannot distinguish among alternative theories, even though the theories have radically different policy implications. The implication of this result for Phillips curve trade-offs is that useful inferences can be made but only by estimating reduced forms over different sub-periods or countries across which policy rules differed systematically. As of 1976, Sargent thought that not much of this type of work had been done, so that little was known.

Second, further empirical work seemed to be consistent with the view that a permanent trade-off exists. King and Watson (1994) study the US Phillips correlations and Phillips trade-offs in a bivariate time-series analysis. They use the unit root (I(1)) inflation process to get around the Sargent (1976) problem (see Fisher and Seater 1993; King and Watson 1997, for details), and estimate structural models to interpret the data and compute the long-run trade-offs and sacrifice ratios (cost of lowering inflation) associated with each model. Except for the extreme case of a real business cycle model, they find long-run trade-offs between inflation and unemployment.

The same conclusion is reached by Akerlof et al. (1996), but for a different reason. They report evidence of permanent downward wage stickiness, which implies a long-run trade-off. This evidence comes from four sources: ethnographic surveys, Bureau of Labor Studies data on the distribution of wage changes in manufacturing establishments, union settlements (in both the United States and Canada), and the authors' own survey of individuals in the Washington DC area. The authors' were aware that Panel Study of Income Dynamics (PSID) data showed evidence of extensive downward wage flexibility, but argue that individual reporting errors are large, and when corrected for using data from the Current Population Survey, downward rigidity is present. The presence of downward wage rigidity would constitute a serious challenge to the natural rate hypothesis – the neutrality of the unemployment rate with respect to the anticipated inflation rate. And not surprisingly, much work has been done to check the conclusion reached by Akerlof, Dickens and Perry. Parkin (2000) summarizes this work, which concludes that the money wage rate is not downwardly rigid and that the appearance of downward rigidity results from three sources of bias; measurement error, rounding error and long-term contracts. Controlling and correcting for these sources of bias points towards wage flexibility. Clearly more work is needed to settle this issue.

The third challenge to monetary neutrality comes from a series of papers by Barro (1977, 1978) and Mishkin (1982a, b). Decomposing money growth into anticipated and unanticipated components, Barro reports that only unanticipated money growth influences unemployment and real GDP and (as predicted) both anticipated and unanticipated money growth influences the price level. Mishkin shows that Barro's estimation procedure, while providing consistent parameter estimates, delivers incorrect standard errors. When Mishkin replicates Barro's exercises with valid tests, he rejects the restrictions implied by neutrality. (He does not reject the restrictions implied by rationality.)

The literature just reviewed deals with the consequences of anticipated inflation and not its causes. Questions concerning causality are more naturally addressed in the context of an investigation of unanticipated inflation.

Unanticipated Inflation

It is not possible to analyse unanticipated inflation in isolation, independently of other aspects of aggregate economic performance. Fluctuations (at the business cycle frequency) in the general level of economic activity and in inflation, though far from perfectly correlated, share some common features. There is, for example, a general positive correlation between inflation and real income (or equivalently, a negative correlation between inflation and unemployment). There is also a positive correlation between money and income as well as between the velocity of circulation of money and income.

The 'stylized facts' about the business cycle (shared by all economies) raise difficult questions about cause and effect. Of the four variables – the price level, real output, the money supply and the velocity of circulation – which, if any, is the prime mover? Do fluctuations in the growth rate of the money supply cause fluctuations in the other variables? Do autonomous movements in the price level, perhaps stemming from wage-push pressure, initiate the fluctuations in money, velocity and output? Does the business cycle have its origin in real factors that initiate fluctuations in output, which in turn lead to induced fluctuations in money supply growth, inflation and velocity?

At one level questions such as these are statistical and are capable of being investigated using econometric methods that detect causality, such as those proposed by Granger (1969). Studies based on such methods have not, however, delivered decisive results.

Most investigations of the possible causes of inflation have sought to understand the phenomenon by identifying the sources of inflation and studying the transmission mechanism whereby those sources are translated into variations in the rate of inflation and in other economic aggregates. This approach is one which seeks to understand both inflation and the business cycle as an integrated phenomenon.

There are three broad classes of theories that have been proposed for understanding the unanticipated and cyclical aspects of inflation. The first of these stems from the work of Keynes (1936) and emphasizes both price stickiness and the potential for autonomous movements in prices. On this view, the normal state of affairs would be one in which wages and prices are relatively sticky, responding only gradually to aggregate demand shocks. Shocks to aggregate demand arise from a variety of sources. One possibility is that autonomous fluctuations in investment produce fluctuations in aggregate demand. Other possible sources of aggregate demand fluctuations are fluctuations in wealth and interest rates which in turn are induced by fluctuations in the growth rate of the money supply. Fluctuations in wealth and interest rates can induce fluctuations in investment and consumption. All of these potential sources of variation in aggregate demand lead to cycles in both output and the price level. Initially, a change in demand will have bigger output effects than price-level effects, but eventually prices and wages will adjust to reflect fully the change in aggregate demand. The resulting co-movements in output and prices will be positively, though not strongly, correlated.

From time to time this normal state of affairs is disturbed by autonomous price shocks. The most commonly hypothesized source of price shocks is wage-push. It is suggested that, at times of substantial industrial or social unrest, movements in the level of money wages will act as a type of social safety mechanism. The idea that wage-push results from sociological phenomena was particularly popular amongst economists in the UK in the early 1970s (see, in particular, Balogh 1970; Jones 1972; Wiles 1973; Hicks 1974. By the time the first oil shock occurred (late 1973), 'wage-push' gave way to 'oil-push' as the most commonly identified source of autonomous movement in inflation.

When autonomous movements in the price level occur, the phenomenon that came to be known as 'stagflation' quickly follows. The autonomous price rise raises the inflation rate and lowers output (raising unemployment). If the higher unemployment and lower output induces an increase in the growth rate of the money supply, then even further price-level rises occur.

This traditional version of the Keynesian theory of inflation and the business cycle, together with some of the sociological embellishments that have been briefly reviewed above, is very thoroughly explained and elaborated in Laidler and Parkin (1975).

More recent and sophisticated versions of the Keynesian theory of cycles and inflation may be found in papers by Fischer (1977), Phelps and Taylor (1977), and Taylor (1979, 1980). The essence of these 'New Keynesian' theories is the existence of long-term contractual arrangements in labour markets. Such arrangements result in wages, the major element of costs, being predetermined. This stickiness of wages and costs

results in a stickiness of prices, even if the expectations of prices that form the basis for the long-term labour market contracts are formed rationally.

A second approach to understanding cyclical fluctuations is one based on incomplete contemporaneous information about aggregate demand. This approach, sometimes called the 'New Classical Theory', was first suggested in the early 1970s by Lucas (1972, 1973). The approach is broadly consistent with the Keynesian mechanism of aggregate demand determination but proposes an alternative theory of aggregate supply. Individual economic agents are assumed to operate in informationally isolated 'islands' and to be incapable of distinguishing relative from absolute price level changes. The resulting confusion causes them to respond to absolute price changes as if they were relative price changes. This response results in positive co-movements in output and the price level.

In both the Keynesian and New Classical approaches, the key driving variable generating the cycle – fluctuations in both real output and the inflation rate – is a fluctuating growth rate in the money supply. This is not to deny that other things might, from time to time, shock the economy. Rather, it is a proposition about the major ongoing source of cyclical variation. Within both of the theories, positive co-movements of velocity are explained by appealing to the idea that to some degree the cycle itself is forecastable. To the extent that it is, higher rates of inflation at the cyclical peak will in part be anticipated and, therefore, reacted to. It is always efficient to reduce money holdings when the opportunity cost of holding money increases. Higher expected inflation rates, leading to higher nominal interest rates, induce such economizing and are, therefore, the major source of procyclical fluctuations in velocity.

A third approach to understanding aggregate fluctuations denies the primacy of variations in the money supply growth rate, or in any other sources of aggregate demand fluctuation in generating the cycle. This approach, known as 'real business cycle theory', has yet to gain a major following but has, in recent years, begun to spawn a growing and important literature (see, in particular, King and Plosser 1984; Kydland and Prescott 1982; Long and Plosser 1983; Nelson and Plosser 1982). Though differing in details, the essential proposition of the new real business cycle theories is that aggregate fluctuations emanate from technological shocks to the aggregate production function or, in some versions, from sector-specific shocks and from the interactions between sectors of the economy – although a large literature has now incorporated Calvo (1983) price stickiness or monopolistic competition.

Technological shocks that generate fluctuations in full-employment output would, other things equal, generate negative co-movements in prices, and, presumably, to the extent that such movements were forecastable, countercyclical movements in velocity. Since such co-movements do not occur, it seems as if the real cycle theories are in substantial trouble. King and Plosser (1984) address this problem directly by proposing that technological shocks which affect real output induce responses in money and credit that accommodate – indeed over-accommodate – the real fluctuations. Thus, when there is a positive shock to aggregate supply, this induces an even bigger rise in the total volume of money and credit and, therefore, induces procyclical co-movements in money, prices and output. To the extent that these are forecastable, economizing on real balances generates procyclical velocity.

There is not, at the present time, any definitive and systematic evidence capable of disposing convincingly of any of these three alternative approaches; nor is there any overwhelming evidence suggesting that any of them is clearly in the lead.

Inflation in Open Economies

The alternative approaches to understanding inflation that have been reviewed so far have (implicitly) examined inflation in a closed economy. Most practical concerns about inflation arise in individual countries which are open economies.

The international trade and international capital market transactions undertaken by such countries have an important bearing on their inflation performance. Also, the foreign-exchange rate regime – fixed or flexible – has an important influence upon a country's inflation performance. It was during the period of rapidly accelerating inflation in the 1970s that open economy theories and the international transmission mechanism gained in prominence (see Parkin and Zis 1976a, b).

The main feature of the analysis of inflation in an open economy is the emphasis on the limited potency of domestic monetary policy under fixed exchange rates. In a country, or more interestingly in a world, operating on fixed exchange rates, individual countries' monetary policies have no effect on the country's rate of inflation. Instead, monetary policy influences the country's balance of payments. In such a world, inflation is a world phenomenon, not a national phenomenon. It is the growth rate of the world money supply that determines the world average rate of inflation. Theorizing along this line had, in fact, made good progress even as early as the middle of the 18th century at the hands of David Hume (1752). It was rediscovered and popularized in the 1960s and early 1970s by Mundell (1971) and Johnson (1973).

The rediscovery of David Hume's analysis provided interesting insights into the resurgence of world inflation at the end of the 1960s. An attempt on the part of the United States to finance its Great Society programme and the Vietnam War with limited tax increases and with an increase in the growth rate of the money supply – with an increase in the inflation tax – became the engine of an inflation that engulfed the entire fixed exchange-rate world.

Understanding the international generation and transmission of inflation in a flexible exchange rate world, such as that which had emerged by the mid-1970s, is still far from settled. At the centre of the problem of understanding inflation is the problem of understanding the determination of foreign exchange rates. Large and rapid movements in foreign-exchange rates are seen as having a potentially powerful and rapid effect on domestic price levels. The forces that determine exchange rates are still, however, far from well understood. Viewing the foreign exchange rate as following a random walk is as precise as any structural theories of the exchange rate that have so far been proposed and tested.

Despite the absence of a convincing theory of inflation in an open economy, the effects of policy coordination (or its absence) have been studied. A central question addressed by Oudiz and Sachs (1984) and Obstfeld and Rogoff (2002) is whether unilateral national monetary policy rules are inferior to international monetary coordination. The answer is that they are not.

Positive Theories of Central Bank Behaviour

Recent developments in understanding inflation have been dominated by the rational expectations revolution and the related and more far-reaching revolution that has uses rigorous dynamic general equilibrium analysis. Some of the implications of that revolution have been discussed above and have been to strengthen and refine the theories of inflation that emphasize fluctuations in the growth rate of the money supply as the principal source of fluctuations in inflation and other economic aggregates.

The rational expectations hypothesis holds that expectations are formed by making predictions of future inflation on the basis of the mechanisms that generate actual inflation. If inflation is indeed caused by rapid monetary expansion, then forecasting future inflation is the same thing as forecasting future monetary policy. But monetary policy itself emerges from an ill-understood political process. In most countries the task of formulating monetary policy has been delegated to a central bank. Yet, in determining monetary policy, central banks are often influenced by the economic and political environment in which they operate and must also take account of the consequences of their actions for the behaviour of the economy as a whole.

In order to understand the inflationary process, with people forming expectations rationally, it becomes necessary to understand the

policymaking mechanisms and the forces that generate varying monetary growth rates. The first serious analysis of this problem was that by Kydland and Prescott (1977) and the problem has been investigated more recently by Barro and Gordon (1983a, b) and Cukierman (1992). In the models proposed by these writers, a central bank's goal is to achieve an optimal combination of inflation and unemployment. Lower inflation and lower unemployment are seen by the central bank as desirable objectives. The bank is constrained, however, by a short-run trade-off between inflation and unemployment – a trade-off arising from the considerations described above. A surprise rise in inflation would produce a cut in unemployment while a surprise drop in inflation would produce a rise in unemployment. The precise way in which the short-run trade-off between inflation and unemployment constrains the central bank depends on the expectations of private agents concerning the bank's behaviour. A central bank that can credibly precommit to a particular rule about inflation – perhaps a zero-inflation rule – would be a bank that could engender rational expectations of zero inflation. It would be optimal for such a bank to in fact precommit to a zero rate of inflation and then deliver that rate.

The ability to precommit and with credibility seems to require some mechanism for binding the central bank that does not have a readily identifiable counterpart in the real world. Central banks are, in fact, free to pursue whatever policies they wish at their discretion. Since this fact is known to all private economic agents, it will be rational for them to take it into account when forming expectations about central bank behaviour. The equilibrium that results in this case will be such as to ensure that the actual inflation rate chosen by the bank is one that removes any temptation for the bank to depart from that rate and further exploit the short-run trade-off. Put differently, the inflation rate chosen will be the best available at the natural rate of unemployment. Only in such a situation would the central bank have no further temptation to attempt to exploit the short-run trade-off. Thus without the ability to precommit to a fixed (and presumably zero) rate of inflation, a central bank will end up delivering a higher rate of inflation than that which is socially desirable.

One feature of the positive theories of inflation developed by Kydland–Prescott and Barro–Gordon that some people find disquieting is the time inconsistency. (In game theory language, the equilibrium concept is Nash rather than sub-game perfection.) Attempts to develop positive analyses that do not have this feature have been based on reputation. One such approach, in Barro and Gordon (1983a), uses the so-called 'trigger strategy' model of reputation suggested by James Friedman (1971). A model is proposed in which the central bank would be punished if it delivered too high a rate of inflation and in which it takes time to restore the bank's reputation. In equilibrium, the bank never does inflate at a rate that requires the punishment to be inflicted.

An alternative approach by Barro (1986) uses the reputation analysis developed by Kreps and Wilson (1982). In this model there are two potential 'types' of central banker, one that likes inflation and one that dislikes it. The inflationary central banker has an incentive to masquerade as a non-inflationary type in order to induce low inflation expectations. By inducing low inflation expectations, the inflationary central bank will, at some point, be able to exploit those low expectations and produce a surprise inflation; it will do this by following initially a strategy of inflating at exactly the same rate as would be chosen by a non-inflationary central bank. At some later point it will pursue a mixed strategy – a strategy analogous to choosing an inflation rate by drawing numbers from an urn. Once this mixed strategy has resulted in a high rate of inflation, the inflationary central banker is revealed, and expectations about inflation as well as actual inflation will rise.

Another feature of the Kydland–Prescott and Barro–Gordon models that is objectionable is that the central bank targets an unemployment rate below the natural rate. If it were to target the natural rate, there is no tension between its inflation and real goals. Cukierman overcomes this objection by replacing the symmetric loss function of the standard model with an asymmetric loss function: the central bank weighs positive

deviations from the natural unemployment rate more heavily than deviations below the natural rate.

Backus and Driffill (1985) and Cukierman (see Cukierman 1992, ch. 3), have suggested another modification to the standard model: the possible interactions between labour unions (working as a unified wage-setting institution) and the central banks. In this case, inflation (and money supply growth) is determined as the outcome of a game between the central bank and the economy-wide labour union.

Empirical tests of the alternative positive theories of central bank behaviour have been conducted by Ruge-Murcia (2003) and by Cukierman and Gerlach (2003). Ruge-Murcia uses US time-series data and rejects the Barro–Gordon formulation but does not reject the Cukierman asymmetric loss function formulation. Cukierman and Gerlach use data for 22 OECD countries and reach a similar conclusion.

Other recent developments in understanding central bank behaviour arise from the normative analysis of monetary policy to achieve an inflation target, and it is convenient to discuss this topic in the context of inflation policy below. But, before that, it is convenient to consider the links between monetary policy and fiscal policy.

Inflation and Fiscal Policy

A further consequence of the rational expectations and dynamic general equilibrium revolutions has been to force attention back to the connection between fiscal and monetary policy. The simple accounting fact that government expenditure must be financed, either by taxation, by borrowing or by money creation, implies that any analysis of the determination of money growth must at the same time make consistent propositions about fiscal policy and deficit financing. Of course, variations in the growth rate of interest-bearing debt can provide a good deal of insulation of money growth from the deficit. Nevertheless, large and persistent deficits may give rise to rational expectations of future money growth, even in the face of currently firm monetary policies. Sargent and Wallace (1981) have shown that, if the fiscal authority is the prime mover and follows taxation and spending policies that are independent of monetary policy, then, essentially, inflation and, ultimately, money growth are fiscal phenomena. Whether these findings are of practical importance is a matter of some controversy. Sargent (1982), studying the ends of four big inflations, has argued that adjustments in fiscal policy have been crucial to ending inflation. By implication, the emergence of a large and apparently uncontrolled deficit would be seen as the origin of serious inflation. Work by Dornbusch and Fisher (1986) offers a different interpretation, however, placing major importance on the behaviour of the foreign exchange rate.

The link between fiscal policy and inflation is most complete in Woodford's (1995) fiscal theory of the price level. Because the quantity of money demanded depends on the opportunity cost of holding money, which in turn depends on the rational expectation of the inflation rate, there is a large number (infinite) of equilibrium price level paths. The standard (mostly unstated) approach rules out all the purely speculative equilibria and selects the unique equilibrium based on the monetary fundamentals. In which the government's choice of how to finance its debt determines the inflation rate. The fiscal theory of the price level rejects this approach and rules out equilibria by the government's selection of its debt financing regime. As an example, Kocherlakota and Phelan (1999) show that, with a policy of constant taxes and constant money, the fiscal theory predicts that a one-time cut in the quantity of money generates a speculative hyperinflation (in contrast to the standard model prediction of a one-time fall in the price level).

Policy Towards Inflation

Analyses of policies towards inflation have changed over the years. Advocacy of gradually slowing down the growth rate of the money supply and advocacy of controls on wages and prices were the most commonly heard policy

suggestions for controlling inflation in the 1960s and early 1970s. Those who saw autonomous wage and price movements as the principal source of inflation saw prices and incomes policies as the major weapon to control it. Those who saw money growth as the source of inflation embraced monetary gradualism as the most obvious cure. A prodigious amount of work attempting to evaluate alternative policies was undertaken, much of which is surveyed by Laidler and Parkin (1975).

As a consequence of the rational expectations and dynamic general equilibrium revolutions, the focus of the policy debate has shifted markedly from that of seeking to manipulate variables such as key wage settlements (the prices-and-incomes policy solution) or the growth rate of the quantity of money (the monetarist solution). Instead, attention has turned to thinking about the way in which different institutional arrangements interact to produce different inflation rates. And the emphasis has shifted from policy as an action to policy as a process or set of rules.

One line of research has examined the consequences of alternative monetary systems, including the adoption of alternative forms of commodity money (see, in particular, 'Conference on Alternative Monetary Standards' 1983). Another research direction has been the investigation of targeting nominal income growth as a means of conquering and avoiding inflation (Tobin 1983; Taylor 1985).

But the idea that has attracted most attention both in the research community and among central banks is the use of a monetary policy rule that seeks to achieve either an inflation rate target or a price level target. The study of inflation or price level targeting has both a positive and a normative dimension and sometimes the two are not explicitly distinguished.

Svensson (1999) has provided a nice distinction between what he calls 'instrument rules' and 'targeting rules' for monetary policy. In the context of inflation targeting (and that is the context of most of the recent literature on monetary policy) an instrument rule specifies how the policy instrument responds to the current state of the economy. The current state can include current forecasts of future variables. A targeting rule, in contrast, states that the policy instrument shall be set at the level that makes for forecast inflation rate equal the inflation target.

The policy instrument that features in instrument rules is either the overnight interest rate on inter-bank loans or the monetary base. Woodford (2003) provides the authoritative account and discussion of the interest rate instrument rule and shows that such a rule can, in principle, deliver low and stable inflation provided that it incorporates the 'Taylor principle', which states that the interest rate must change in the same direction as a change in the inflation rate but by more than the change in the inflation rate (Taylor 1993, 1999).

McCallum (1988) has explored the use of a monetary base rule and compared the robustness of interest rate and monetary base rules.

It is a curious fact about the models that explore the use of an interest rate rule that money plays either no role or no essential role in the inflation process. The models in which money plays no role are typically specified as reduced forms in which inflation is generated by expected inflation and the output gap; the output gap responds to the real interest rate, which equals the nominal interest rate set by the central bank minus the inflation rate; and expectations are rational. Other models are specified at a deeper structural level with consumers maximizing intertemporal utility of consumption and leisure and monopolistically competitive firms setting prices according to a Calvo (1983) formula.

In some models, money enters through a 'shopping time' function (King and Wolman 1996). But whether present in the model or not, money plays no essential role in the inflation process. This fact is emphasized in Woodford (2003) by his exploration of the cashless economy and is seen as a virtue because it might provide insights on inflation in a future economy when technological change has driven money, as we know it, out of existence.

It is also a curious fact that inflation targeting amounts to targeting a variable whose value cannot be influenced by a central bank's current actions until well beyond the bank's forecast horizon. It is the long and variable lags in the response

of inflation (and output) to monetary policy that led Friedman to his original advocacy of a money stock growth rate target.

The evolution of inflation over the coming years will provide valuable evidence on both the inflation process, the currently out-of-favour monetarist ideas, and the wisdom of the current policy regimes.

Conclusion

Macroeconomics in general, and the theory of inflation in particular, is in a fluid state. The foregoing has attempted to review that state and provide a picture of the path that we have taken in getting to it. We have broad agreement on the facts to be explained and broad agreement on the behaviour of nominal variables (for given real variables) in an inflationary economy in which the path of inflation is anticipated. We also have broad agreement that fully anticipated inflations, though in many theoretical models capable of generating non-neutralities, are nevertheless to a good approximation neutral. Beyond that there is little in the way of firm knowledge. We have a variety of models of macroeconomics and inflation, and many clear theoretical results. We do not have much, however, in the way of solidly based rejections of any of the available models. Uncertainty surrounds both the issue of the impulse (or impulses) that generate inflation and other fluctuations and the issue of the propagation mechanisms that translate those impulses into movements in output and the price level.

See Also

- ▶ Central Bank Independence
- ▶ Cost-Push Inflation
- ▶ Demand-Pull Inflation
- ▶ Hyperinflation
- ▶ Inflation Dynamics
- ▶ Inflation Expectations
- ▶ Inflation Measurement
- ▶ Inflation Targeting
- ▶ Neutrality of Money

Bibliography

Akerlof, G.A., W.T. Dickens, and G.L. Perry. 1996. The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1996 (1): 1–76.

Backus, D., and J. Driffill. 1985. Rational expectations and policy credibility following a change in regime. *Review of Economic Studies* 3: 211–221.

Balogh, T. 1970. *Labour and inflation*. London: Fabian Society.

Barro, R.J. 1977. Unanticipated money growth and unemployment in the United States. *American Economic Review* 67: 101–115.

Barro, R.J. 1978. Unanticipated money, output, and the price level in the United States. *Journal of Political Economy* 86: 549–580.

Barro, R.J. 1986. Reputation in a model of monetary policy with incomplete information. *Journal of Monetary Economics* 17: 3–20.

Barro, R.J. 1997. *Determinants of economic growth: A cross-country empirical study*. Cambridge, MA: MIT Press.

Barro, R.J., and D.B. Gordon. 1983a. Rules, discretion and reputation: A model of monetary policy. *Journal of Monetary Economics* 12: 101–121.

Barro, R.J., and D.B. Gordon. 1983b. A positive theory of monetary policy in a natural rate model. *Journal of Political Economy* 91: 589–610.

Bernanke, B.S., and M. Woodford. 2004. *The inflation-targeting debate*. Chicago: University of Chicago Press.

Bronfenbrenner, M., and F.D. Holzman. 1963. A survey of inflation theory. *American Economic Review* 53: 593–661.

Calvo, G.A. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Clower, R.W. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6: 1–8.

Conference on alternative monetary standards. 1983. *Journal of Monetary Economics* 12(1).

Cukierman, A. 1992. *Central bank strategy, credibility, and independence*. Cambridge, MA: MIT Press.

Cukierman, A., and S. Gerlach. 2003. The inflation bias revisited: Theory and some international evidence. *The Manchester School* 71: 541–565.

Dornbusch, R., and S. Fischer. 1986. Stopping hyperinflations past and present. *Weltwirtschaftliches Archiv* 122: 1–47.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85: 191–205.

Fischer, S. 1979. Anticipations and the non-neutrality of money. *Journal of Political Economy* 87: 228–252.

Fisher, M., and J. Seater. 1993. Long run neutrality and superneutrality in an ARIMA framework. *American Economic Review* 83: 402–415.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.
Friedman, M. 1970. *The counter-revolution in monetary theory*. London: Institute of Economic Affairs.
Friedman, J.W. 1971. A non-cooperative equilibrium for supergames. *Review of Economic Studies* 38: 1–12.
Granger, C.W.J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37: 424–438.
Hicks, J.R. 1974. *The crisis in Keynesian economics*. Oxford: Blackwell.
Hume, D. 1752. Of money; of interest; of the balance of trade. *Political discourses;* reprinted in *Three essays: Moral, political and literary.* London: Oxford University Press, 1963.
Johnson, H.G. 1963. A survey of theories of inflation. *Indian Economic Review* 6 (4): 29–69.
Johnson, H.G. 1973. Secular inflation and the international monetary system. *Journal of Money, Credit, and Banking* 5 (1, Part II): 509–520.
Jones, A. 1972. *The new inflation: The politics of prices and incomes*. London: Penguin Books and André Deutsch.
Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan.
King, R.G., and C.I. Plosser. 1984. Money, credit, and prices in a real business cycle model. *American Economic Review* 74: 363–380.
King, R.J., and M.W. Watson. 1994. The post-war U.S. Phillips curve: A revisionist econometric history. *Carnegie-Rochester Conference Series on Public Policy* 41: 157–219.
King, R.J., and M.W. Watson. 1997. Testing long-run neutrality. *Federal Reserve Bank of Richmond Quarterly Review* 83: 69–101.
King, R.G., and A.L. Wolman. 1996. Inflation targeting in a St. Louis model of the twenty-first century. *Federal Reserve Bank of St. Louis Review* 78 (3): 83–107.
Kocherlakota, N., and C. Phelan. 1999. Explaining the fiscal theory of the price level. *Federal Reserve Bank of Minneapolis Quarterly Review* 23 (4): 14–23.
Kormendi, R.C., and P.G. Meguire. 1985. Macroeconomic determinants of growth: Cross-country evidence. *Journal of Monetary Economics* 16: 141–163.
Kreps, D., and R. Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27: 253–279.
Kydland, F.E., and E.C. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85: 473–491.
Kydland, F.E., and E.C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.
Laidler, D., and M. Parkin. 1975. Inflation: A survey. *Economic Journal* 85: 741–809.
Long, J.B., and C.I. Plosser. 1983. Real business cycles. *Journal of Political Economy* 91: 39–69.
Lucas, R.E. Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
Lucas, R.E. Jr. 1973. Some international evidence on output-inflation tradeoffs. *American Economic Review* 63: 326–334.
McCallum, B.T. 1988. Robustness properties of a rule for monetary policy. *Carnegie-Rochester Conference Series on Public Policy* 29: 173–203.
Mishkin, F.S. 1982a. Does anticipated monetary policy matter? An econometric investigation. *Journal of Political Economy* 90: 22–51.
Mishkin, F.S. 1982b. Does anticipated aggregate demand policy matter? Further econometric results. *American Economic Review* 72: 788–802.
Mundell, R.A. 1963. Inflation and real interest. *Journal of Political Economy* 71: 280–283.
Mundell, R.A. 1965. Growth, stability and inflationary finance. *Journal of Political Economy* 73: 97–109.
Mundell, R.A. 1971. *Monetary theory: Inflation, interest and growth in the world economy*. Pacific Palisades: Goodyear Publishing Co.
Nelson, C.R., and C.I. Plosser. 1982. Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10: 139–162.
Obstfeld, M., and K. Rogoff. 2002. Global implications of self-oriented national monetary rules. *Quarterly Journal of Economics* 117: 503–535.
Oudiz, G., and J. Sachs. 1984. Macroeconomic policy coordination among the industrial economies. *Brookings Papers on Economic Activity* 1984 (1): 1–76.
Parkin, M. 2000. *What have we learned about price stability? Price stability and the long-run target for monetary policy.* Ottawa: Bank of Canada.
Parkin, M., and G. Zis. 1976a. *Inflation in open economies*. Manchester: Manchester University Press.
Parkin, M., and G. Zis. 1976b. *Inflation in the world economy*. Manchester: Manchester University Press.
Phelps, E.S. 1968. Money, wage dynamics, and labor market equilibrium. *Journal of Political Economy* 76: 678–711.
Phelps, E.S., and J.B. Taylor. 1977. Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85: 163–190.
Ruge-Murcia, F.J. 2003. Does the Barro–Gordon model explain the behavior of US inflation? A reexamination of the empirical evidence. *Journal of Monetary Economics* 50: 1375–1390.
Sala-i-Martin, X., G. Doppelhoffer, and R.I. Miller. 2004. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94: 813–835.
Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482.
Sargent, T.J. 1976. The observational equivalence of natural and unnatural rate theories of macroeconomics. *Journal of Political Economy* 84: 631–640.
Sargent, T.J. 1982. The ends of four big inflations. In *Inflation: Causes and effects*, ed. R.E. Hall. Chicago: University of Chicago Press.
Sargent, T.J., and N. Wallace. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis, Quarterly Review* 5 (3): 1–17.

I

Schwartz, A.J. 1973. Secular price change in historical perspective. *Journal of Money, Credit, and Banking* 5 (1, Part II): 243–269.
Sidrauski, M. 1967. Inflation and economic growth. *Journal of Political Economy* 75: 796–810.
Stockman, A.C. 1981. Anticipated inflation and the capital stock in a cash-inadvance economy. *Journal of Monetary Economics* 8: 387–393.
Svensson, L. 1999. Inflation targeting as a monetary policy rule. *Journal of Monetary Economics* 43: 607–654.
Taylor, J.B. 1979. Staggered wage setting in a macro model. *American Economic Review, Papers and Proceedings* 69 (2): 108–113.
Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.
Taylor, J.B. 1985. What would nominal GNP targeting do to the business cycle? *Carnegie-Rochester Conference Series on Public Policy* 22: 61–84.
Taylor, J.B. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.
Taylor, J.B., ed. 1999. *Monetary policy rules*. Chicago: University of Chicago Press.
Tobin, J. 1965. Money and economic growth. *Econometrica* 33: 671–684.
Tobin, J. 1983. Monetary policy: Rules, targets and shocks. *Journal of Money, Credit, and Banking* 15: 506–518.
Townsend, R. 1980. Models of money with spatially separated agents. In *Models of monetary economies*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.
Wallace, N. 1980. The overlapping generations model of fiat money. In *Models of monetary economies*, ed. J.-H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.
Wiles, P. 1973. Cost inflation and the state of economic theory. *Economic Journal* 83: 377–398.
Woodford, M. 1995. Price level determinacy without control of a monetary aggregate. *Carnegie-Rochester Conference Series on Public Policy* 43: 1–46.
Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

Inflation Accounting

G. Whittington

The adjustment of the accounts of business enterprises to reflect the consequences of inflation has been the subject of considerable theoretical controversy and practical experimentation in recent years, as a result of historically high inflation rates. The traditional valuation basis of accounts is historical cost, and the gap between historical values and current values tends to be widened by inflation. Furthermore, historical cost accounting does not reflect the gain on borrowing, which arises from having a liability which is fixed in nominal monetary units, in a period of inflation, or its counterpart, the loss on holding money, or assets denominated in nominal monetary units (such as trade debts).

Two apparently competing systems have been proposed to deal with the inflation accounting problem. The first is Current Purchasing Power Accounting, CPP. This applies general price indices to his torical cost in order to allow for the decline in the value of money due to inflation. The second system is current value accounting, which revalues assets and liabilities at their current values, or, alternatively, restates historical cost by reference to a specific price index relating to the specific asset type, rather than a general price index. The current value accounting system found most commonly in practice, and featuring most prominently in the theoretical debate, is current cost accounting, CCA, which revalues assets at current cost, that is, replacement cost (typically) or recoverable amount (the higher of net selling price or net present value of future services to its present owners), whichever is the lower. A third system, which combines current valuation with CPP adjustments, and therefore eliminates the need to choose between CPP and CCA, is the real terms system, RT.

The principles of the three systems may be illustrated by a simple numerical example. Suppose an asset is bought at time *O* for £10,000 when the retail price index is 100. By time *1*, the current cost of the asset is £15,000, by which time, the retail price index is 120. CPP accounting would restate the historical cost (£10,000 at time *O form of physical cap*) as £12,000 (= £10,000 × 120/100) at time *1*, because it would take 1.2 time *1* £s to buy the equivalent of 1 time *0* £. On the other hand, the proprietor's capital also needs to be restated in an exactly equivalent manner, so no gain or loss is shown by CPP, unless the asset was financed by borrowing in nominal money units (which would not require

restatement, leading to a gain on borrowing) or had a value fixed in nominal money units (which would also not require restatement, leading to a loss on holding monetary assets). Current value accounting, on the other hand, would restate the asset at its current market value, £15,000 at time *1*. If we retained the unindexed money capital maintenance convention, this would lead to the recognition of a gain of £5000 (= £15,000–£10,000). Alternatively, if we used a form of physical capital maintenance convention, such as is commonly found in CCA systems, we would restate capital also by reference to the specific price change of the physical assets, so no gain would be shown. Real Terms, RT, accounting, on the other hand, would restate the asset at current value (£15,000 at time *1*) but would restate the proprietor's capital by reference to the retail price index (£12,000 at time *1*), showing a 'real gain' of £3000 (= £15,000–£12,000), which is the amount by which the asset has appreciated in excess of the fictitious element due to inflation. There are many detailed variations on the simple systems illustrated here. These are explained and illustrated in Whittington (1983).

The CPP system deals only with the effects of general price level changes and ignores relative changes in the prices of specific assets. It is thus a pure inflation accounting system which would be adequate if all prices changed in the same proportion. The system was developed in Germany during the hyperinflation of the early 1920s. It was developed further and introduced into the North American literature by H.W. Sweeney (1936). During the past two decades, similar systems have been adopted by leading Latin American countries (notably Brazil, Chile and Argentina) under the pressure of very high inflation rates. In practice, CPP seems to be adopted only when pure inflation is seen as an important and urgent problem.

The CCA system owes its origins to the replacement cost accounting systems proposed by American, Dutch and German writers in the first three decades of the 20th century. CCA became prominent in the inflation accounting debate in the English-speaking world in the mid-1970s, with the aid of support from government agencies (such as the Securities and Exchange Commission in the USA and the Sandilands Committee in the UK). An account of this 'CCA counter-revolution' (and other aspects of the history of inflation accounting) will be found in Tweedie and Whittington (1984). The probable motivation for government support for CCA was that this system avoids the use of general indices and that any form of general indexation was regarded, at the time, as reinforcing the process of inflation. The obvious strength of CCA is that it attempts to record the assets held and used by the business at their specific current prices (although the precise definition of current cost is a controversial issue), thus measuring the current performance (in the profit and loss account) and state of the business (in the balance sheet) in terms of economic opportunities currently available in the market place. The weakness of the system is in its treatment of assets and liabilities which are of fixed nominal money value. The two capital maintenance concepts which are naturally associated with CCA are physical capital maintenance (as in the Sandilands Committee's operating profit measure) and nominal money capital maintenance (as in the Sandilands Committee's statement of total gains). Neither of these is capable, in its simplest form, of reflecting the gain on borrowing or loss on holding money which occurs in a period of inflation: indeed, some supporters of CCA would deny that such gains and losses occur. In order to capture these effects, recent British CCA systems (as in the Accounting Standard SSAP16 1980) have adopted the gearing adjustment and the monetary working capital adjustment. These attempt to capture the gain on borrowing and loss on holding money, respectively, by using specific rather than general price indices. They have proved difficult to implement in practice as well as being difficult to justify theoretically, and are currently under review. They have been proposed but not implemented in a number of other countries (Australia, Canada, Germany, New Zealand, and Sweden). A clear account of the debate surrounding the introduction of CCA in the United Kingdom will be found in Kennedy (1978).

The RT system owes its origins to the work of Sweeney (1936), who pointed out that stabilization of the monetary unit by the use of general indices could be applied to any valuation base. The term CPP is normally restricted to stabilization of a historical cost base, so we use the term RT for stabilization of a current cost or other current value base. Since a CCA system already records assets at their current values, which are denominated in current currency units, the stabilization of such a system to convert it to an RT system requires no further indexation of asset values. The proprietor's capital is, however, adjusted by a general index, so that initial capital is maintained in terms of real purchasing power before a profit or gain is recognized. Thus, the RT system recognizes only real gains on assets (as illustrated earlier): it also recognizes the real gain on borrowing and loss on holding monetary assets in a period of inflation. This system was developed in considerable detail by Edwards and Bell (1961), who showed that it was possible, within the RT framework, to calculate CCA operating profit and then, by adding real gains and losses on holding assets and liabilities (which Edwards and Bell describe as 'real holding gains' or losses) to derive a final total of real profit or gain (which they describe as 'real business profit').

In many ways, the RT system seems to be a logical and consistent means of recognizing the effects of general inflation in eroding the purchasing power of proprietor's capital, while also recognizing the effects of changing individual prices on the value or cost of the specific assets held and used by the business. It thus deals with the problems dealt with by both CPP and CCA, while avoiding the weaknesses of these two systems. There are strong elements of the RT system in the current USA standard on accounting for changing prices (FAS33), and it may be the system which will ultimately prevail in practice. Its slow emergence has much to do with the fact that CPP and CCA have been espoused by groups which see their particular interests being served by one of these systems; for example, professional accountants tend to be attracted by the relative objectivity (and therefore lower risk of professional liability for error) of CPP, which avoids subjective estimates of current values.

Finally, it should be noted that this essay has dealt only with business accounting. Inflation accounting is also an important problem in national accounts. The traditional adjustments for price changes are replacement cost of capital consumed and the elimination of stock appreciation (see Stone and Stone 1977). This is analogous to CCA adjustment of business accounts. National accounts are also often restated in constant price terms, but this is a crude transformation rather than CPP or RT adjustment which would require restatement of opening capital figures to reveal real holding gains and losses (including those on monetary items) in various sectors. This issue is explored in Godley and Cripps (1983).

See Also

- ▶ Accounting and Economics
- ▶ Historical Cost Accounting

Bibliography

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

FAS33. 1979. *Financial reporting and changing prices*, Statement of Financial Accounting Standards No.33. Stamford: Financial Accounting Standards Board.

Godley, W., and F. Cripps. 1983. *Macroeconomics*. Oxford: Oxford University Press.

Kennedy, C. 1978. Inflation accounting: Retrospect and prospect. *Cambridge Economic Policy Review*, (4): 58–64.

Sandilands Report. 1975. Inflation accounting: Report of the inflation accounting committee under the chairmanship of F.E.P. Sandilands, Cmnd. 6225. London: HMSO, 1975.

SSAP16. 1980. *Current cost accounting*, Statement of Standard Accounting Practice No.16. London: Accounting Standards Committee.

Stone, J.R.N., and G. Stone. 1977. *National income and expenditure*, 10th ed. Cambridge: Bowes & Bowes.

Sweeney, H.W. 1936. *Stabilized accounting*. New York: Harper.

Tweedie, D.P., and G. Whittington. 1984. *The debate on inflation accounting*. Cambridge: Cambridge University Press.

Whittington, G. 1983. *Inflation accounting: An introduction to the debate*. Cambridge: Cambridge University Press.

Inflation and Growth

John Cornwall

Although the relationship between inflation and economic growth has interested economists for some time, the nature of this association is still not well understood. Early discussions deliberated on the relative merits of a rising compared to a falling price level on profits, confidence, investment and other macro variables as these affected the growth of the economy, especially productivity. No noticeable consensus emerged from these deliberations. In more recent times, in particular the post World War II period up until the early 1970s, the historical record gives ambiguous if not misleading clues. For example, cross-country comparisons of rates of inflation and productivity growth in the developed capitalist economies reveal virtually no association between the two. And if the period of rapid growth of productivity of the 1950s and 1960s is compared with the period of stagnation since the early 1970s, over time a negative correlation between inflation and productivity growth is found in each of the economies. The rise in inflation rates is associated with a slowdown in productivity growth.

The Political Economy of Inflation and Growth

Conventional economic theory is equally inconclusive on the causal connection between the two. However, since the early 1970s, activist government intervention in the various economies has introduced a connecting link, resulting in a definite causal connection between inflation and growth that is likely to persist for some time to come. As a result, a correct understanding of this relationship involves a conceptual framework that is broader than that assumed by conventional economic theory. The causal relationship between inflation and growth must be seen as a problem in political economy, for it is the response of governments to inflation, both actual and predicted, that has and will determine the ultimate impact of inflation on the growth of productivity and output.

To put the matter in its simplest form, inflation leads to slow growth or stagnation because in those countries in which inflation cannot be brought under control by other means (e.g. an incomes policy) governments respond by implementing restrictive aggregate demand policies. Such responses lead, as they have since the early 1970s, to high rates of unemployment and low rates of capacity utilization, investment and productivity growth.

Taking the analysis one step further, in studying the mechanism by which inflation leads to stagnation under existing institutions it is useful to divide the developed capitalist economies into two groups. First, there are economies that experience accelerating rates of inflation under conditions of sustained high employment. To put the matter differently, there are countries in which the non-accelerating inflation rate of unemployment (NAIRU) is greater than the rate of unemployment at which all unemployment is voluntary.

No successful incomes policy can be implemented that would allow involuntary unemployment to be reduced to a minimum without the strong demand conditions leading to accelerating rates of inflation. As a result these countries will adopt restrictive aggregate demand policies in order to increase unemployment enough to reduce the rate of inflation. The fear that stimulative fiscal policies will lead to greater budget deficits and the fear of increased power of labour under full employment conditions, partly because it is believed that each causes inflation rates to accelerate, will reinforce this trend towards restrictive aggregate demand policies.

These economies can be said to suffer from an inflationary bias (Cornwall, 1983, ch. 6). Because of the policy response to this bias, inflation (or even the fear of inflation) will lead to high rates of unemployment and low rates of growth of productivity, that is, economic stagnation.

In contrast there is a second group of economies which, because of favourable institutional arrangements, could achieve full employment

without accelerating rates of inflation *if restrictive aggregate demand policies were not adopted by the first group of economies*. These countries would be likely to adopt full employment policies if restrictive policies were not in effect elsewhere. But when they are, this group of economies is also forced to pursue restrictive aggregate demand policies but for quite different reasons than the first group. However, the effect of policy on the growth of output and productivity is the same; it will be greatly reduced.

Pluralist Economies

Thus the first step in understanding the relation between inflation and productivity growth is a recognition that the simultaneous achievement of full employment and non-accelerating rates of inflation is not an automatic feature of capitalist economies. Moreover any failure to achieve these goals is not to be attributed to a failure of the authorities to follow some monetary or fiscal rule. Instead, the failure of an economy to handle inflationary pressures while maintaining full employment must be attributed to existing institutional and political arrangements. These make the coordination of wage and price settings in individual markets with the national goal of overall wage and price stability impossible.

These institutions can be said to act as constraints limiting the number and kinds of policy instruments available to the authorities for combating inflation. Going further, since the authorities in these economies respond to accelerating inflation by creating whatever unemployment is politically tolerable in an attempt to reduce inflation, the use of aggregate demand policies as an instrument for realizing desirable employment goals is, therefore, severely constrained by an inflationary bias. The authorities in these economies can be expected to pursue stagnationist policies under existing institutional and political arrangements.

The relation between inflation, the political response to inflation, and growth just described is similar to that seen by Kalecki (1977). However, it is more accurate to limit the kind of 'political theory of the business cycle' foreseen by Kalecki to a special group of capitalist economies which will be referred to as 'pluralist' economies. Pluralist economies share certain features in common. Governments play an essentially passive role in governing, primarily reacting to demands by special interest groups; there is a widespread belief among the powerful economic and political groups that an invisible hand or system of countervailing power exists that guarantees some kind of social optimum; the industrial relations system can be characterized as adversarial; and decision-making within the trade-union movement is decentralized.

The countries today that suffer from an inflationary bias and whose institutional features most clearly coincide with those just mentioned are the developed English-speaking countries, particularly Canada, the United Kingdom and the United States. Very likely, other countries with somewhat different institutions suffer from an inflationary bias, for example, France and Italy, and for the purposes at hand could be included in this group (Barber and McCallum, 1982; Crouch, 1984; McCallum, 1983).

Corporatist Economies

There is a second group of economies which will be referred to as 'corporatist economies'.

Corporatist economies are characterized by a tradition of state intervention in the economy, a high degree of cooperation and collaboration between the major economic groups in policy formation, a disbelief in invisible hands, and a system of industrial relations that can be described as cooperative. Primarily because of these institutions, these economies have been able in one form or another to implement relatively successful voluntary incomes policies in the past. Inflation has not been eliminated to be sure but has been kept at rates lower than would likely have resulted had union and managament been unwilling to cooperate with government in the interests of achieving wage and price stability. More certainly, as Table 1 reveals, economies with these characteristics and with powerful trade union movements as well, for

Inflation and Growth, Table 1 Annual average rates of unemployment (U) and inflation ($\dot{p}$), 1963–73 for selected capitalist economies

	U^a	$\dot{p}$		U^a	$\dot{p}$		U^a	$\dot{p}$
Austria	1.7%	4.2%	Italy	5.2%	4.0%	Switzerland	0.0[b]	4.5%
Canada	4.8	4.6	Japan	1.2	6.2	Sweden	2.0	4.9
France	2.0	4.7	Netherlands	1.2	5.5	United Kingdom	3.0	5.3
Germany	0.8	3.6	Norway	1.7	5.3	United States	4.5	3.6

[a]1965–1973
[b]National definition
Source: OECD, *Economic Outlook*, Paris, various issues; and OECD, *Labour Force Statistics*, Paris, various issues

example, Austria and Sweden, have been able to reduce unemployment to extremely low levels without experiencing inflation rates much higher (if higher) than those in the pluralist economies.

Unfortunately given the high degree of economic interdependence between economies, most of those economies best able to contain inflation at full employment can no longer do so when the pluralist economies adopt restrictive policies. The economic importance of the pluralist bloc in the world economy forces restrictive policies on the second group of countries. Their importance guarantees that by restricting aggregate demand in their own countries, depressed conditions in the pluralist countries will be exported to the others in the form of a decrease in demand for their exports. Furthermore any attempt by any of the corporatist economies to offset declining exports through stimulative demand policies will lead to current account deficit that cannot be sustained through continuous borrowing (Thirlwall and Hussain, 1982). As a result, the full employment goal must be sacrificed.

Basic to this argument is the assumption that in the face of depressed demand conditions in the pluralist bloc, any corporatist economy acting on its own is not able to offset the adverse effects of a full employment policy on its payments position through an exchange rate policy. Unfortunately changes in the exchange rate are not sufficient to induce the kind of expenditure switching needed to bring the current account of the corporatist economies more or less into balance at full employment. These economies can be said to be limited or constrained in their use of aggregate demand policies for attaining full employment because of a payment constraint.

It is useful for pedagogical reasons to divide the capitalist world into two mutually exclusive groups, pluralist and corporatist. With this simplification in mind, the stagnating capitalist economies fall into one or the other of two groups: those in which restrictive demand policies are employed out of a fear of inflation and those in which a fear of payments problems at full employment caused by the restrictive policies of the first group leads to the same policies. The causal mechanism at work today, whereby inflation (or merely the fear of inflation) in one group of countries leads to worldwide stagnation, now becomes clear. As long as the pluralist group restricts aggregate demand because of an inflationary bias, less than full employment conditions are forced upon the rest of the world. As a result, an inflationary bias in the pluralist group, that is, a tendency for inflation rates to accelerate at or before full employment, leads not just to breakdown in those countries but to worldwide stagnation.

The Failure of Conventional Policies

As just argued, worldwide stagnation can be attributed to an inflationary bias in a group of key countries. Seen in another way, the difficulties or sources of stagnation can be traced to a failure of the traditional policy instruments, that is, monetary, fiscal and exchange rate policies, to work successfully in realizing full employment, price stability and external balance. Underlying this failure are certain structural and institutional changes that develop over a prolonged period of full employment such as the quarter of a century following World War II. Simply put, in

democratic capitalist societies the rising affluence attributable to a long period of full employment is accompanied by the extension of the welfare state. This greatly increases the relative power of labour. As a result wages (and prices) are no longer determined primarily by the traditional market forces of demand and supply (Hicks, 1974; Okun, 1981; Scitovsky, 1978). This makes aggregate demand policy a highly inefficient means of fighting inflation. While wages and prices may respond eventually if restrictive policies are pushed far enough, the quantity effects on output and employment are substantial and immediate and persist while the policy is in effect. Futhermore, any implementation of an expansionary demand policy following a 'successful' restrictive policy that has brought down inflation rates will merely bring back the inflation in the pluralist economies. Restrictive policies whose aim is to permanently reduce inflation will fail in these countries because they fail to attack the sources of the inflationary bias.

Increased affluence and greater labour power also contribute to the ineffectiveness of exchange rate policy. First, consider that the trend in international trade has been increasingly towards the more highly fabricated goods that are desired for their non-price qualities, for example, design, durability, reliability, delivery dates, etc. This trend can, to a large extent, be attributed to affluence. It results in a downward trend in price elasticities of traded goods making it increasingly unlikely that the Marshall–Lerner conditions will be satisfied. When they are not, devaluation leads to a worsening of the trade deficit, other things being equal.

Second, the successful use of the exchange rate as an instrument for relieving a payments constraint is severely compromised by the existence of real wage resistance. A cheapening of exports relative to imports following devaluation likely leads to a decline in real wages. Under full employment conditions labour will have a strong incentive to press for higher money wage increases in an effort to protect their real wages. The resulting wage–price spiral can lead to the real exchange rate returning to its previous level. Like the inflationary bias, this difficulty arises out of the increased power of labour under full employment conditions and the affluence full employment brings.

Real wage resistance can be a real problem even in corporatist economies that may have had success in the use of income policies in the past. In earlier times the incentive for acceptance by labour of a voluntary incomes policy was provided by a promise of full employment and the rising real wages that full employment encourages. Unfortunately a devaluation of the currency, forced upon the authorities in their pursuit of full employment by restrictive policies in the pluralist bloc, may lead to non-compliance with the incomes policy.

If the reduction in real wages can be limited, real wage resistance might be avoided. However, the international interdependence of capital markets can and has lead to situations in which the local authorities have little control over the magnitude of the actual depreciation of the exchange rate. A deliberate devaluation generates fears of accelerated inflation in the minds of managers of exceedingly large and mobile capital funds. This leads to large withdrawal of funds from the country, a further depreciation of the currency, greater fears of inflation, etc. As the experience of several countries in the recent past make clear, governments are soon forced to reverse their employment policies in order to protect the exchange rate.

Conclusions

In order to break the causal chain leading from inflation to restrictive policy to stagnation that prevails under modern capitalist conditions, major structural-institutional changes are required. Most important are changes that would relieve the pluralist economies of their inflationary bias. A different conception of the role of the state in the economy, the development of a cooperative industrial relations system and possibly of centralized collective bargaining would be extremely helpful because these changes increase the possibility that a successful incomes policy could be implemented. The benefits of its success for the rest of the world are apparent.

A second programme for recovery is more limited in that it is restricted to the corporatist bloc. This would take the form of the corporatist economies adopting coordinated trade and lending policies that discriminate against the pluralist group, in order to ease possible payments difficulties from full employment policies.

There are other possibilities involving one or more countries, but, however much they differ in detail, they share one thing in common: all require radical and basic structural changes in key economic and political institutions. Without these adaptations, the present political economy of inflation and stagnation will continue indefinitely and will be worldwide.

See Also

- ▶ Forced Saving
- ▶ Inflation
- ▶ Stagflation
- ▶ Supply Shocks in Macroeconomics

Bibliography

Barber, C., and J. McCallum. 1982. *Controlling inflation; learning from experience in Canada, Europe and Japan*. Ottawa: Canadian Institute for Economic Policy.

Cornwall, J. 1983. *The conditions for economic recovery: A Post-Keynesian analysis*. Oxford: Blackwell.

Crouch, C. 1984. The conditions for trade-union wage restraint. In *The politics of inflation and economic stagnation*, ed. L. Lindberg and C. Maier. Washington, DC: Brookings.

Hicks, J. 1974. *The crisis in Keynesian economics*. New York: Basic Books.

Kalecki, M. 1977. Political aspects of full employment. In *Selected essays on the dynamics of the capitalist economy 1933–1970*, ed. M. Kalecki. Cambridge: Cambridge University Press.

McCallum, J. 1983. Inflation and social consensus in the seventies. *Economic Journal* 93(December): 784–805.

Okun, A. 1981. *Prices and quantities: A macroeconomic analysis*. Washington, DC: Brookings.

Scitovsky, T. 1978. Market power and inflation. *Economica* 45(August): 221–233.

Thirlwall, A.P., and M. Hussain. 1982. The balance of payments constraint, capital flows and growth rate differences between developing countries. *Oxford Economic Papers* 34(3): 198–510.

Inflation Dynamics

Timothy Cogley

Abstract

There have been a number of changes in monetary policy rules in the United States and UK since the early 1960s. The Lucas critique says that this should induce changes in the equilibrium law of motion. This article summarizes reduced-form evidence on the evolving law of motion for inflation in the USA and the UK. Since the 1970s, inflation has become lower on average, less volatile and less persistent. There is also less uncertainty about the central bank's long-run target for inflation.

Keywords

Bank of England; Bretton Woods system; Central-bank independence; Commitment; European Monetary System (EMU); Federal Reserve System; Fixed exchange rates; Greenspan, A.; Inflation; Inflation dynamics; Inflation gap; Inflation targeting; Inflation volatility; Lucas critique; Monetary policy rules; New Keynesian macroeconomics; Output gap; Phillips curve; Prediction-error variance; Sticky price models; Supply shocks; Taylor rule; Thatcher, M.; Trend inflation; Vector autoregressions; Volcker, P

JEL Classifications

E3

As with other macroeconomic data series, economists have long been interested in the dynamics of inflation. This series is of particular interest because of its relationship to alternative theories of aggregate fluctuations and associated effects of alternative monetary policies. Indeed, for understanding the behaviour of inflation, it is important to take into account the alternative monetary policies that were in force during the period being studied. Shifts in monetary policy rules alter the

fundamentals that drive inflation and therefore also alter its dynamic properties. Accordingly, one must distinguish inflation variation arising within a stable monetary regime from movements that follow from shifts in policy rules.

For the United States, Taylor's (1993) rule is often used to describe Federal Reserve behaviour. Although his rule was originally intended as a normative proposal, Taylor and others soon discovered that it closely approximated Federal Reserve behaviour during the Volcker–Greenspan era (1979–2006). Shortly thereafter, a number of economists began to explore whether the Taylor rule also described Fed behaviour prior to that (for example, Clarida et al. 2000; Lubik and Shorfheide 2004). They found that it did, although with different coefficients. Prior to Volcker's term as chairman, the Fed reacted more strongly to fluctuations in output and was less sensitive to movements in inflation. In fact, although the Fed increased the nominal funds rate when inflation rose, it increased the funds rate by less than one for one, so that the real funds rate actually declined, thus amplifying the initial movement in inflation. According to Clarida, Gali and Gertler and Lubik and Schorfheide, this was an important factor behind the high and volatile inflation of the 1970s.

Important changes have also occurred in UK monetary policy. After the Second World War, the Bank of England at first operated under the Bretton Woods system of fixed exchange rates; the breakdown and float in the early 1970s was followed by attempts to re-establish fixed exchange rates in the 1980s, the decision to opt out of EMU after the foreign-exchange crisis of 1992, the adoption of inflation targeting in 1992, and finally central-bank independence from the Treasury in 1997. The last two steps in particular altered the way the Bank of England conducts monetary policy, with the Bank now placing a higher priority on controlling inflation.

The Lucas critique says that a change in a government policy rule should alter the equilibrium law of motion for endogenous variables. In this article I summarize research on changes in the dynamic properties of inflation since the early 1960s. I focus on the USA and the UK because they are the two economies studied most extensively in the literature.

Evolving Inflation Dynamics in the USA and the UK

I use a vector autoregression (VAR) to summarize the dynamic properties of inflation. For a historical period during which government and private decision rules are unchanged, one can estimate a time-invariant VAR. Here I am concerned about a period with changing monetary policy rules, however, so VAR parameters must be allowed to vary, in accordance with the Lucas critique. Consequently, much of the literature on evolving inflation dynamics studies Bayesian VARs with drifting conditional mean and variance parameters; for example, Benati and Mumtaz (2006), Canova and Gambetti (2006), Cogley and Sargent (2001, 2005a, b), Cogley et al. (2005), and Primiceri (2005a).

Cogley and Sargent (2005a, b) estimate VARs of the form

$$y_t = X_t'\theta_t + \mathrm{E}_t, \tag{1}$$

where y_t is a vector consisting of inflation, a short-term nominal interest rate, and a real activity variable such as unemployment or GDP. The right-hand variables X_t consist of a constant plus lags of y_t, and the conditional mean parameters θ_t evolve as a driftless random walk subject to reflecting barriers. The driftless random walk assumption makes θ_t vary as

$$\theta_t = \theta_{t-1} + v_t, \tag{2}$$

where v_t is $NID(0, Q)$. The reflecting barriers prevent θ_t from wandering into the region of the parameter space where the system has explosive autoregressive roots. This representation puts a unit root in inflation, because the reflecting barriers do not restrict drift in the VAR intercepts, but it prohibits more than one unit root in inflation.

The VAR innovations ε_t are assumed to be conditionally normal with mean zero and drifting variance

$$R_t = B^{-1} H_t B^{-1\prime}, \tag{3}$$

where H_t is diagonal and B is lower triangular:

$$H_t = \begin{pmatrix} h_{1t} & 0 & 0 \\ 0 & h_{2t} & 0 \\ 0 & 0 & h_{3t} \end{pmatrix}, \tag{4}$$

$$B = \begin{pmatrix} 1 & 0 & 0 \\ \beta_{21} & 1 & 0 \\ \beta_{31} & \beta_{32} & 1 \end{pmatrix}. \tag{5}$$

The diagonal elements of H_t are univariate stochastic volatilities that evolve as driftless, geometric random walks:

$$\ln h_{it} = \ln h_{it-1} + \sigma_i \eta_{it}. \tag{6}$$

The random-walk assumption is designed to fit permanent shifts in innovation variances such as those associated with the 'Great Moderation' in the USA (McConnell and Perez Quiros 2000). This formulation allows time-varying correlations among the VAR innovations, and it guarantees that R_t is positive definite. Primiceri (2005a) extends the model to allow for drifting covariances as well, $R_t = B_t^{-1} H_t B_t^{-1\prime}$.

The results reported below illustrate various aspects of the Bayesian posterior distribution for this model. Readers who are interested in the technical details should consult the original sources.

Trend Inflation

Figure 1 depicts trend inflation in the USA and the UK. Following Beveridge and Nelson (1981), I define trend inflation in terms of long-horizon forecasts. At date t, trend inflation is the level at which inflation is expected to settle after the transient variation dies out,

$$\overline{\pi}_t = \lim_{j \to \infty} E_t \pi_{t+j}. \tag{7}$$

To approximate $\overline{\pi}_t$, write the VAR in companion form as

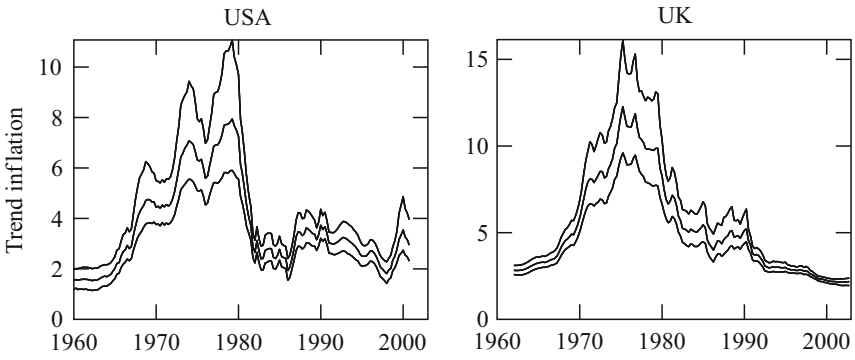
$$z_t = \mu_t + A_t z_{t-1} + \varepsilon_{zt}, \tag{8}$$

where μ_t contains the VAR intercepts and A_t the autoregressive parameters. Then trend inflation $\overline{\pi}_t$ can be approximated by the long-horizon VAR forecast,

$$\overline{\pi}_t \doteq \mu_t (I - A_t)^{-1}. \tag{9}$$

The figure portrays the posterior median of $\overline{\pi}_t$ at each date along with the interquartile range. (All the figures are based on the author's calculations.)

The median estimate of trend inflation was a bit below two per cent in the USA in the early 1960s, and it was just shy of three per cent in the UK. It increased sharply in both countries in the late 1960s and early 1970s and peaked in the mid- to late 1970s. In the UK, a peak of 12 per cent was



Inflation Dynamics, Fig. 1 Trend inflation in the USA and the UK (*Sources*: Federal Reserve Economic Database (USA), Bank of England (UK))

reached in 1975, and $\overline{\pi}_t$ remained in double digits until 1980. In the USA, trend inflation ratcheted to 4.5 percent in 1970, then to seven per cent in 1974, and finally to almost eight per cent in 1979. In the early 1980s, Paul Volcker's disinflation in the USA and Margaret Thatcher's monetarist experiment in the UK brought $\overline{\pi}_t$ quickly back down to more tolerable levels. In the USA, trend inflation has fluctuated around 2–3.5 per cent since 1985. In the UK, $\overline{\pi}_t$ settled in the neighbourhood of 2.5 per cent when the Bank of England adopted an explicit inflation target in 1992, and then declined gradually to around two per cent after 1997.

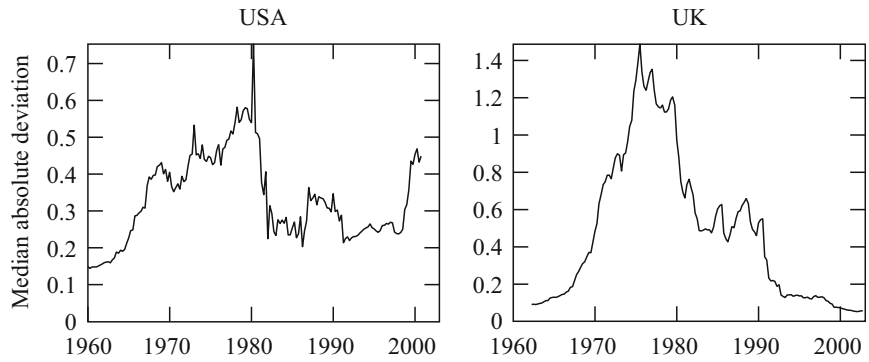
Measures of uncertainty about trend inflation also rise and then fall. The inter-quartile range for $\overline{\pi}_t$ is narrow at the beginning of the sample, widens substantially in the middle, and then narrows again at the end. Thus, the 1970s was not only a decade when trend inflation was high, but also a time of substantial uncertainty about its exact value. For example, in the USA, when the median estimate of $\overline{\pi}_t$ peaked at eight per cent, there was a fifty-fifty chance that it was somewhere outside the interval 5.75–12 per cent. Similarly, when the median estimate of UK trend inflation peaked at 12 per cent, there was a fifty-fifty chance that it could exceed 16 per cent or fall short of 12 per cent.

The inter-quartile range quantifies level uncertainty, that is, uncertainty about where $\overline{\pi}_t$ is at a particular date. One can also quantify how rapidly trend inflation drifts. Figure 2 portrays the median absolute deviation for $\Delta\overline{\pi}_t$. (I report this statistic instead of the standard deviation because of outliers.) Not only is there a lot of uncertainty about the level of $\overline{\pi}_t$ in the 1970s, but $\overline{\pi}_t$ was also drifting more rapidly then. The rate of drift fell considerably in both countries in the early 1980s, and in the UK it also declined sharply after 1992. Stock and Watson (2005) were the first to report a result like this in the context of an unobserved-components model of US inflation. The result also holds for our drifting-parameter VARs.

Uncertainty about $\overline{\pi}_t$ presumably reflects uncertainty about the central bank's long-run target for inflation, or doubt about its commitment to the target, or both. For the UK, it is interesting to note how the inter-quartile range narrowed and the rate of drift declined after the adoption of inflation targeting in 1992. In contrast, the inter-quartile range for the USA was about as wide in the 1990s as in the 1980s, and its width was also comparable to that of the UK for the 1980s. Similarly, the rate of drift in US trend inflation was considerably higher at the end of the sample than in the UK. The difference, of course, is that the USA has not adopted a formal inflation target. Taken at face value, these figures illustrate how an explicit inflation target can anchor long-run inflation expectations.

Inflation Gap Variability

Next I turn to changes in inflation volatility. To a first-order approximation, trend inflation is a random walk, which means that inflation itself has



Inflation Dynamics, Fig. 2 Median absolute deviation for $\Delta\overline{\pi}_t$ (*Sources*: Federal Reserve Economic Database (USA), Bank of England (UK))

infinite unconditional variance. Here I focus instead on the volatility of de-trended inflation, $\pi_t - \overline{\pi}_t$. A central bank that behaves as if minimizing an undiscounted quadratic loss function will adjust its policy instrument so that inflation eventually converges to its target. Thus, I interpret $\overline{\pi}_t$ as a measure of target inflation and $\pi_t - \overline{\pi}_t$ as the deviation from the target or 'inflation gap'. At each date in the sample, I approximate its instantaneous standard deviation as

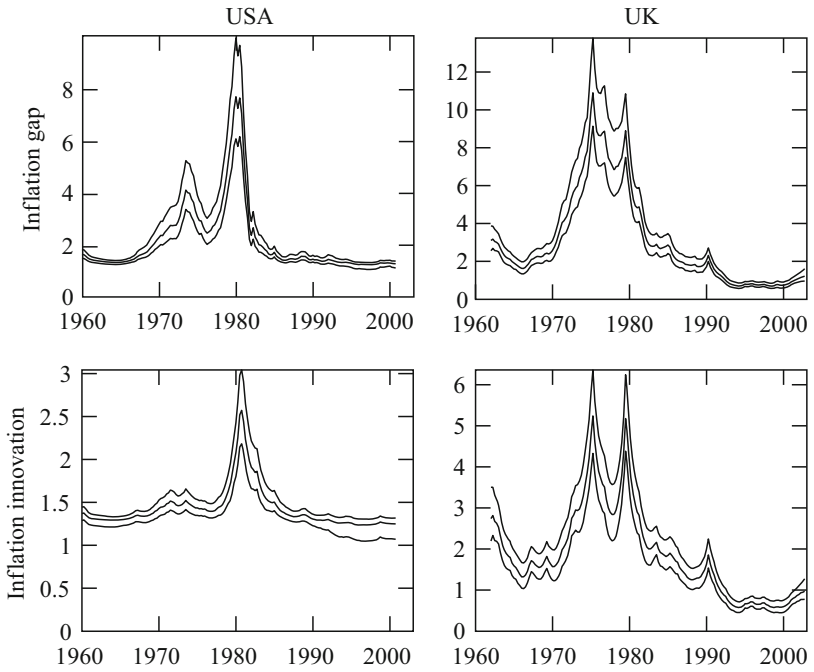
$$\sigma_t(\pi_t - \overline{\pi}_t) = \left[\sum_{j=0}^{\infty} e_\pi' A_t^j R_t A_t^{j\prime} e_\pi\right]^{1/2}, \qquad (10)$$

where e_π' is a selection vector that picks out π_t from the vector z_t. The top row of Fig. 3 portrays the evolution of the posterior median and interquartile range for σ_t.

Along with Fig. 1, Fig. 3 depicts a positive correlation between inflation gap volatility and trend inflation. Inflation volatility was low at the beginning and ends of the samples when trend inflation was low, and it was high during the great inflation of the 1970s. Thus, not only has average inflation declined since the 1970s, but so has inflation gap volatility. Moreover, this did not come at the expense of an increase in output or unemployment volatility, which follow much the same trajectory as for inflation. Sticky price models often predict a trade-off between the variability of inflation and real variables, but that trade-off is not apparent here. Instead we see a simultaneous decline in both after 1980.

Whether the greater stability of inflation and output is the result of better policy or better luck (that is, smaller shocks) is the subject of much current research. The absence of a volatility trade-off suggests that better luck is a promising candidate, however, for smaller shocks would deliver a simultaneous decline in both in standard models.

Inflation can be volatile either because shocks are volatile or because they are persistent. Thus, we can drill down by examining innovation variances and measures of persistence. In the bottom row of Fig. 3, I report the standard deviation of one-step ahead VAR prediction errors for inflation. For the UK, the pattern is the same as in the other figures:



Inflation Dynamics, Fig. 3 Standard deviation of inflation gaps and inflation innovations (*Sources*: Federal Reserve Economic Database (USA), Bank of England (UK))

prediction errors were large in magnitude during the 1970s and smaller before and after. For the USA, there was only a slight increase during much of the 1970s, but a sharp spike during the brief window in 1979–80 when the Federal Reserve was targeting monetary aggregates.

To the extent that monetary policy affects inflation with long and variable lags, these pictures also hint that good luck in the form of smaller shocks is part of the story. But the movements in innovation variances do not necessarily disprove the bad-policy story, for better policy can take the form of smaller policy shocks. It is also conceivable that better policy could damp the impact of non-policy shocks. Perhaps more importantly, if policy in the 1970s was so bad that sunspots affected equilibrium outcomes, then better policy could eliminate one of the shocks altogether (that is, the sunspot), and that would reduce the VAR prediction error variance for inflation and other variables.

Stock and Watson (2005) also report a decline in the prediction-error variance after the great inflation, but they point out another sense in which inflation has simultaneously become harder to forecast. Consider the R^2 statistic for the VAR forecast of inflation,

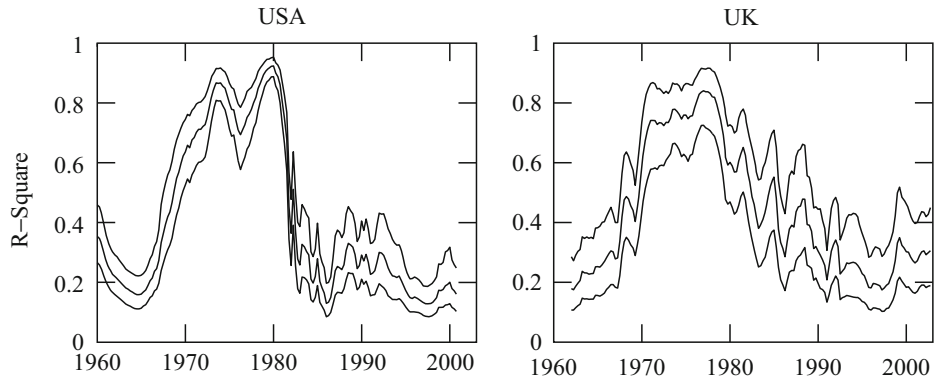
$$1 - R_t^2 = \frac{\sigma_t^2(\pi_t - E_{t-1}\pi_t)}{\sigma_t^2(\pi_t - \overline{\pi}_t)}. \qquad (11)$$

The numerator is the VAR innovation variance shown in the bottom row of Fig. 3, and the denominator is the total variance depicted in the top row. Since both terms of the ratio decline after the great inflation, it is not obvious whether the R^2 statistic has increased or decreased. One-step ahead prediction errors are smaller after 1980, but so is the total amount of transient variation that one hopes to predict.

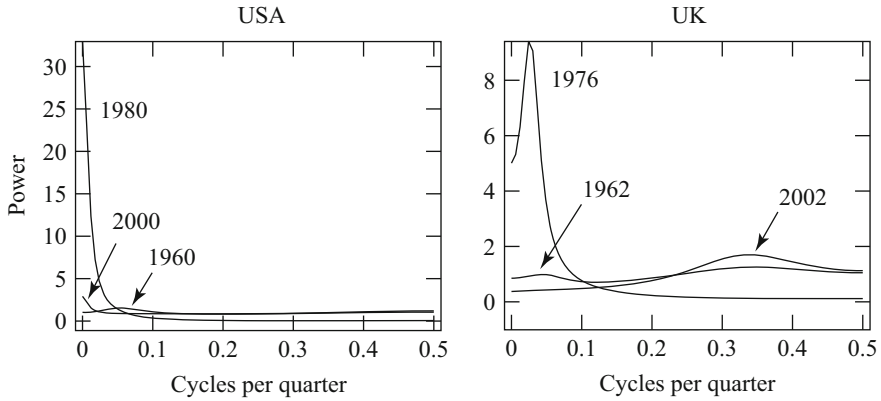
As shown in Fig. 4, for our time-varying VARs the denominator actually falls by more, so that the R^2 for inflation declines. Furthermore, this decline is statistically significant. For example, for the USA the posterior probability of a decline in R^2 between 1980 and 2000 is 0.998. Stock and Watson report a similar finding for an unobserved-components model of inflation. Thus, inflation has become both more and less predictable: inflation forecast errors are smaller in absolute value, but they account for a larger proportion of inflation-gap variability.

This means the inflation gap has become less autocorrelated and also less crosscorrelated with lags of other macroeconomic variables. In other words, it is closer to a martingale-difference variate. In Fig. 5, I summarize changes in inflation gap persistence by graphing its normalized spectrum,

$$g_{\pi\pi}(\omega, t) = \frac{2\pi f_{\pi\pi}(\omega, t)}{\int_{-\pi}^{\pi} f_{\pi\pi}(\omega, t) d\omega} \qquad (12)$$



Inflation Dynamics, Fig. 4 Predictability of the inflation gap (*Sources*: Federal Reserve Economic Database (USA), Bank of England (UK))



Inflation Dynamics, Fig. 5 Normalized spectrum for the inflation gap (*Sources*: Federal Reserve Economic Database (USA), Bank of England (UK))

In the numerator, $f_{\pi\pi}(\omega, t)$ represents the instantaneous power spectrum,

$$f_{\pi\pi}(\omega, t) = e_{\pi}'(I - A_t e^{-i\omega})^{-1} \frac{R_t}{2\pi} \left(I - A_t e^{i\omega}\right)^{-1\prime} e_{\pi}. \tag{13}$$

The denominator is the instantaneous variance of $\pi_t - \overline{\pi}_t$. Thus, $g_{\pi\pi}(\omega, t)$ measures autocorrelation rather than autocovariance. I also multiply by 2π so that the units are easy to interpret. In these units, a white noise variate has $g_{\pi\pi}(\omega, t) = 1$ at all frequencies. Relative to that benchmark, excess power at low frequencies represents positive autocorrelation, and excess power at high frequencies signifies negative autocorrelation. If the ordinate at frequency zero is less than 1, then the price level is partially mean reverting (Cochrane 1988).

In the early 1960s, the spectrum was relatively flat in both countries, and the inflation gap was not far from being white noise. The gap became more persistent by the mid- to late 1970s, however, with power concentrated at frequencies of eight years per cycle or longer. This signifies the presence of substantial transient fluctuations in inflation. Evidently, the monetary authorities were permitting inflation fluctuations to go unchecked for years at a time, only gradually bringing $\overline{\pi}_t$ back toward $\overline{\pi}_t$. These policies were reversed after the early 1980s, and by the end of the sample the spectrum had again become relatively flat. Thus, we also see a positive correlation between trend inflation and persistence of the inflation gap. Trend inflation was low and the gap was weakly persistent at the beginning and the end of the sample, and they were high and strongly persistent, respectively, in the middle.

For the USA, Cogley and Sargent (2005b) and Primiceri (2005b) explain this association in terms of changing Fed beliefs about the sacrifice ratio. In Cogley and Sargent's model, the central bank wants to reduce inflation in the 1970s, but it wants to move very slowly. Their hypothetical central bank prefers gradualism because it puts some weight on Keynesian Phillips-curve models which at that time predicted intolerable sacrifice ratios – much higher than the predictions of the same models in the 1980s or 1990s. Thus, when inflation was highest, optimal policy called for an extremely gradual adjustment towards the target, making the inflation gap highly persistent.

Cogley and Sargent's story gains credibility when one reviews the analyses of leading policy economists from the late 1970s. For example, Arthur Okun (1978, p. 284) wrote that 'recession will slow inflation, but only at the absurd cost in production of roughly $200 billion per point'. At that time, $200 billion amounted to roughly ten per cent of GDP, and, if we extrapolate Okun's estimate to zero inflation, the total cost amounts to three-quarters of a year's GDP. Like the central bank in Cogley and Sargent's model, Okun recommended gradualism in the 1970s because he thought the cost of aggressive actions would be exorbitant.

This explanation dovetails nicely with the work of Orphanides (2001, 2003), who demonstrates that the Fed overestimated the magnitude of the output gap in the 1970s because it was slow to detect the productivity slowdown. Because the estimated output gap was too big, they also initially exaggerated the amount of disinflation that would ensue. When that disinflation failed to materialize, they became pessimistic about the amount of slack needed to slow inflation, concluding that the sacrifice ratio was bigger than previously thought. Output gap misperceptions are not an element of Cogley and Sargent's model (their hypothetical central bank is better at filtering than the Fed was), but it may be an important part of the bigger picture.

In retrospect, the high estimates of sacrifice ratios in the 1970s may seem excessive because current estimates are quite a bit lower. Indeed, that is probably one reason why central banks now react more strongly to inflation. In any event, what matters for understanding monetary policy in the 1970s was what economists believed then, not what we believe now with the benefit of hindsight.

Finally, it is interesting to contrast the shape of the spectrum for the USA and the UK at the end of the sample. For the UK, the spectrum has a trough at frequency zero and a gentle positive slope, a shape that signifies partial mean reversion in the price level. For the USA, there is still a peak above 1 at frequency zero and a downward sloping spectrum, hence no mean reversion in the price level.

The contrast is noteworthy because it connects with questions about optimal monetary policy. In a textbook version of a dynamic New Keynesian model, the first-order condition for optimal policy is

$$\pi_t = -\frac{\lambda}{\kappa}(x_t - x_t - 1), \qquad (14)$$

where x_t is the output gap and λ and κ are parameters (Woodford 2003, ch. 7). Because x_t is a stationary random variable, the right-hand side is over-differenced, implying that optimal policy induces mean reversion in the price level. Woodford explains that a partially mean-reverting price level is a feature of optimal policy in many versions of the New Keynesian model, because a credible commitment on the part of the central bank to roll back future price increases restrains a firm's incentive to increase its price today. To make that promise credible, the central bank must follow through by taking actions to reverse realized movements in the price level. The end-of-sample UK inflation spectrum implies a partial rollback of the price level, but the US inflation spectrum does not.

Conclusion

During the great inflation of the 1970s, inflation outcomes worsened in many dimensions. Inflation was higher on average, more volatile and more persistent. There was more uncertainty about the central bank's long-run target for inflation and also more uncertainty about where inflation would be one quarter ahead. All of that has been reversed, possibly because of improved monetary policy rules, possibly because we have not experienced the severe adverse supply shocks that central bankers had to contend with in the 1970s. Sorting out the reasons behind the improvement in inflation outcomes is the subject of much ongoing research.

See Also

- ▶ Inflation
- ▶ Inflation Targeting
- ▶ Monetary Policy, History of

Bibliography

Benati, L., and H. Mumtaz. 2006. *The great stability in the UK: God policy or good luck?* Bank of England: Working paper.

Beveridge, S., and C. Nelson. 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics* 7: 151–174.

Canova, F., and L. Gambetti. 2006. Structural changes in the U.S. economy: Bad luck or bad policy? Discussion Paper No. 5457. London: CEPR.

Clarida, R., J. Gali, and M. Gertler. 2000. Monetary policy rules and macroeconomic stability: Evidence and some theory. *Quarterly Journal of Economics* 115: 147–180.

Cochrane, J. 1988. How big is the random walk in GNP? *Journal of Political Economy* 96: 893–920.
Cogley, T., and T. Sargent. 2001. Evolving post-World War II U.S. inflation dynamics. In *NBER macroeconomics annual 16*, ed. B. Bernanke and K. Rogo. Cambridge: MIT Press.
Cogley, T., and T. Sargent. 2005a. Drifts and volatilities: Monetary policies and outcomes in the post World War II U.S. *Review of Economic Dynamics* 8: 262–302.
Cogley, T., and T. Sargent. 2005b. The conquest of U.-S. inflation: Learning and robustness to model uncertainty. *Review of Economic Dynamics* 8: 528–563.
Cogley, T., S. Morozov, and T. Sargent. 2005. Bayesian fan charts for U.K. inflation: Forecasting and sources of uncertainty in an evolving monetary system. *Journal of Economic Dynamics and Control* 29: 1893–1925.
Lubik, T., and F. Shorfheide. 2004. Testing for indeterminacy: An application to U.S. monetary policy. *American Economic Review* 94: 190–217.
McConnell, M., and G. Perez Quiros. 2000. Output fluctuations in the United States: What has changed since the early 1980s? *American Economic Review* 90: 1464–1476.
Okun, A. 1978. Discussion. In *Curing chronic inflation*, ed. A. Okun and G. Perry. Washington, DC: Brookings Institution.
Orphanides, A. 2001. Monetary policy rules based on real-time data. *American Economic Review* 91: 964–985.
Orphanides, A. 2003. The quest for prosperity without inflation. *Journal of Monetary Economics* 50: 633–663.
Primiceri, G. 2005a. Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies* 72: 821–852.
Primiceri, G. 2005b. Why inflation rose and fell: Policymakers' beliefs and U.S. postwar stabilization policy. Working Paper No. 11147. Cambridge, MA: NBER.
Stock, J., and M. Watson. 2005. *Has inflation become harder to forecast?* Mimeo: Harvard University and Princeton University.
Taylor, J. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.
Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.

Inflation Expectations

Bennett T. McCallum

Abstract

Agents' expectations about future values of inflation play an important role in macroeconomic analysis. From a steady-state perspective, higher expected inflation rates induce agents to hold smaller real money balances and, in most models, to hold different amounts of capital. In dynamic analysis, inflationary expectations affect agents' decisions regarding saving and price adjustments, and affect monetary policy behaviour in ways that have become increasingly important. Over time, analysts' treatment of expectations evolved from distributed-lag, adaptive models to rational expectations, a change that had major analytical implications. Analysis of learning behaviour has become more prominent, supplementing or occasionally replacing rational expectations.

Keywords

Bretton Woods system; Commodity money; Distributed lags; E-stability; Fiat money; Fisher, I.; Hyperinflation; Infinite horizons; Inflation; Inflation targeting; Inflationary expectations; Learning in macroeconomics; Leisure; Long-term interest rates; Lump-sum taxes; Monetary policy; Monetary policy rules; Neutrality of money; New Keynesian macroeconomics; New neoclassical synthesis; Phillips curve; Rational expectations; Real business cycles; Real vs nominal interest rates; Recursive least squares learning; Stabilization policy; Steady-state theorizing; Sticky prices; Thornton, H.; Time preference; Transversality condition; Wicksell, K.

JEL Classifications

E3

The concept of inflationary expectations came to the fore in the work of Chicago School economists during the 1950s, with notable contributions including those of Cagan (1956), Bailey (1956), and Friedman (1960, 1969) on hyperinflation experiences, the cost of inflation, and the optimal steady-state inflation rate. The upsurge of inflation experienced in many countries following the 1971–3 demise of the Bretton Woods System led to additional interest, which was increased again by the spread of rational expectations analysis

during the 1970s. Yet another boost in prominence came from the widespread influence of monetary policy strategies based on the notion of inflation targeting, beginning around 1990 and continuing unabated as of 2006. Major developments in technical analysis relating to monetary policy have lent additional interest to inflationary expectations in various ways that are touched upon below.

The following exposition begins by considering ways that inflationary expectations are important in terms of comparative steady-state analysis, pertaining to 'long-run' phenomena, before turning to expectations' role in dynamic analysis that corresponds to cyclical fluctuations. Next, the manner in which expectations are formed is discussed, with emphasis on the rational expectations hypothesis and also on recent attempts to depart in a disciplined manner from the strict rationality requirement. Finally, a brief historical note is included.

Steady-State Effects

From a steady-state perspective it is natural to presume that actual and expected rates of inflation (and other variables) coincide, so it is common to discuss the welfare cost of inflation, super-neutrality, and so on, in terms of actual rather than expected inflation. Most of the allocational effects are, however, attributable in principle to expected rather than realized inflation. Even in an economy in which the real rate of interest is invariant to expected inflation, the nominal interest rate – and therefore the quantity of real money balances held – will be influenced by these expectations. In particular, a relatively high expected inflation rate will induce individuals to hold *(ceteris paribus)* relatively small shares of their wealth in the form of money (which, as the medium of exchange, pays its holders interest at a lower rate – often assumed to be zero – than other assets). Consequently, since reduced real money balances entail reduced quantities of the transaction-facilitating services that are provided by the medium of exchange, agents are required to devote relatively more time and/or resources to the activity of 'shopping', that is, conducting transactions. In addition, the volume of transactions conducted may fall. A reduced level of utility is then the consequence for each individual agent, *ceteris paribus*, of an increased rate of expected inflation. In two classic contributions, Friedman (1960, 1969) argued that, on the assumption that there are virtually no resource costs associated with the creation and management of fiat money, overall efficiency requires a rate of expected inflation that drives the opportunity cost of holding money to zero and thereby satiates agents with the transaction-facilitating services of money. An exposition that extends the argument to models with finite-lived agents and considers the modification needed when lump-sum tax/transfers are not feasible is provided by McCallum (1990).

In many well-articulated models, expected inflation also has steady-state effects on other real macroeconomic variables – that is, money is not 'superneutral' (Barro and Fischer 1976). In models with finite-lived agents, for example, the steady-state real rate of interest will be affected by inflationary expectations and, consequently, so will the per capita stock of capital and rate of consumption. But, even if individuals are modelled as having infinite time horizons and a fixed rate of time preference – features which (together with exogenous growth rates) fully determine the steady- state real rate of interest – capital and consumption per capita will under most specifications depend (though probably weakly) on the expected inflation rate. (The well-known model of Sidrauski 1967, provides an exception, but only because it ignores individuals' desire for leisure.) In sum, the magnitude of inflationary expectations may have significant allocative consequences, even if one neglects the practically important effects of tax schedules that are set in nominal terms. Such allocative effects are in principle operative also at business-cycle frequencies, but the large magnitude of the capital stock/investment ratio in developed countries leads to the presumption that these effects are of quantitative significance only over longer spans of time.

Dynamic Effects

There are three main ways in which expectations of future inflation affect period-byperiod equilibria in dynamic analyses with typical macroeconomic models. First, intertemporal decisions depend significantly on *real* rates of interest, which are nominal rates adjusted for expected inflation. Second, expected inflation rates are important determinants of price-setting behaviour in most models in which there is some form of nominal price stickiness, reflecting a failure of prices to adjust immediately to values that would prevail under full flexibility. Third, monetary policy decisions may be based in substantial part on expected inflation rates, as with the strategy of 'inflation forecast targeting' that has been prominent in recent years (for example, Bernanke et al. 1999; Svensson and Woodford 2005).

With respect to the second of these, there has been much disagreement over the best way to represent departures from full price flexibility. (Indeed, an important school of macroeconomic thought adheres to the real-business-cycle view that it is best to assume full flexibility.) In recent years (for example, 1998–2006) variants of the Calvo (1983) price-adjustment scheme have been most prominent, but over the years specifications due to Lucas (1972a, 1973), Fischer (1977), Taylor (1980), Mankiw and Reis (2002), and others have also attracted significant support. Some of the models advanced in the 1970s imply that any real stimulus resulting from inflation will be smaller, the greater is the extent to which this inflation was previously expected. Indeed, a prominent and important line of thought originated by Friedman (1966, 1968) and Phelps (1967) contends that inflation will provide a stimulus to output and employment (via the so-called Phillips-curve relation) *only* to the extent that it is unexpected. As the validity of that viewpoint – that there is no long-lasting trade-off between unemployment and inflation – is highly relevant for stabilization policy, many attempts were made to conduct statistical tests. The appropriate design of such tests will, of course, depend significantly on the way in which expectations are formed, a matter discussed below.

Rational Expectations

As mentioned above, for steady states it is natural to presume that expected inflation rates will match those actually realized, and virtually all contemporary steady-state theorizing proceeds under that assumption. Analysis of quarter-to-quarter or year-to- year movements requires, however, some more ambitious formulation concerning expectational behaviour. From the time of Cagan's (1956) study of hyperinflations until the mid-1970s, the most widely used hypothesis was that of adaptive expectations – which makes each period's change in the relevant variable proportional to the most recent expectational error – with other autoregressive representations also used to some extent. During the 1970s it became clear, however, that adaptive and other fixed autoregressive specifications permit the occurrence of repeated, systematic expectational *errors.* But, since such errors are costly to the individual agents who make them, standard neoclassical reasoning suggests that it would be analytically fruitful to assume that agents typically eliminate any systematic source of expectational error, subject to available information. This hypothesis of rational expectations was introduced by Muth (1961) and developed in a macroeconomic context by Lucas (1972a, 1973) and Sargent (1973). It met with some initial resistance, perhaps because of a mistaken impression that it implies homogeneity of information and expectations across agents and/or that activist macroeconomic stabilization policy must necessarily be ineffective. Scepticism about agents' cognitive abilities also played a role, probably. But by the end of the 1970s the rational expectations hypothesis – implying that an agent's expectational errors are uncorrelated over time with all elements of his information set – had become dominant in both theoretical and applied macroeconomics.

The early development of techniques for the econometric implementation of rational expectations involved attempts to test the Friedman–Phelps no-trade-off hypothesis mentioned above. Various estimates of the crucial slope parameter attached to the expected-inflation variable in a Phillips-type relationship had been

obtained, during the late 1960s and early 1970s, with econometric procedures relying upon the assumption of adaptive expectations (or fixed autoregressive expectations with lag weights summing to 1.0). Typical estimates of the slope parameter obtained in these studies were in the vicinity of 0.4–0.6, well below the value of unity implied by the Friedman–Phelps theory (for example, Solow 1969). It was shown analytically by Sargent (1971) and Lucas (1972b), however, that the test strategy utilized would not identify the parameter at issue if expectations are in fact formed rationally. Instead, the estimate would tend to equal this parameter value times the sum of lag coefficients in a univariate forecasting equation for the inflation rate, a sum that need not equal the value of 1.0 presumed by the procedure in question. Estimates using similar (quarterly, US) data-sets but taking account of this insight were then found to yield values close to unity (McCallum 1976). The resulting interpretation – that the true parameter value is approximately unity and that expectations are at least approximately rational – subsequently received indirect support from additional estimates presuming fixed autoregressive expectations, as the values obtained rose over time during the 1970s (Gordon 1976). As the univariate autoregressive representation of actual inflation was also changing during this period, with the sum of lag coefficients rising from around 0.5 to nearly 1.0, these findings accorded well with the Sargent–Lucas interpretation of the evidence.

As important implication of the Sargent-Lucas analysis is that, if expectations are in fact rational, one cannot generally measure the 'long-run' effect (that is, comparative steady-state effect) of one variable on another by the sum of coefficients in a distributed-lag relationship. For example, since expected inflation affects interest rates to a different extent from unexpected inflation, the sum of coefficients in a distributed-lag regression of interest on inflation will depend on the stochastic properties of inflation (the variable being forecast) as well as the slope coefficient measuring the effect of expected inflation on interest. To test hypotheses about the latter effect, it is necessary to take some account of the type of process generating the variable being forecast. That this principle continues to obtain when frequency-domain statistical techniques are employed was emphasized by Whiteman (1984) and McCallum (1984), but King and Watson (1992) and Fisher and Seater (1993) developed procedures that can be used for model-free tests if the monetary policy rule in force (over the sample period) is such that it generates unit-root behaviour of the log of the money stock.

During the second half of the 1990s, expectations came to play an increasingly important role in monetary policy analysis as 'New Keynesian' or 'New Neoclassical Synthesis' models, firmly based on optimizing analysis while incorporating sluggish price adjustments, became the norm for researchers in academia and central banks alike (Goodfriend and King 1997; Rotemberg and Woodford 1997; Clarida et al. 1999; Woodford 2003). In these models with forward-looking expectations in the price-adjustment equations, optimal policy requires 'history- dependent' rules (Woodford 2003) that take account of expectations in a manner not recognized in traditional optimal-control analysis. Various developments, including consideration of issues implied by the zero lower bound on nominal interest rates, made policy analysis increasingly a matter of 'managing expectations' (Eggertsson and Woodford 2003). From the perspective of actual policy practice rather than theory, Goodfriend and King (2005) present documentation, based on Federal Open Market Committee transcripts, indicating that as early as November 1979 the committee was using long-term interest rates as an indicator of inflationary expectations, which were being used to help guide the disinflation of 1979–84. This episode has, more recently, come to be widely regarded as a major turning point in the remarkable worldwide reduction in inflationary difficulties that took place between the late 1970s and the early 1990s.

Issues Regarding Expectations

The principle mentioned above, involving distributed-lag coefficient sums, can remain true

even if expectations are not strictly rational. In particular, it will apply if expectations are formed in a manner that reflects full but delayed responsiveness to the properties of the generating system. Expectational behaviour of that type, which might be termed 'asymptotic rationality,' can be expressed analytically by the condition that the unconditional mean of the expectational error process equals zero, a weaker requirement than that the error must be uncorrelated with all information variables available at the time of expectation formation. This less stringent type of partial rationality has not been prominent to date, but may become important eventually. It is not analytically similar, it should be said, to hypotheses involving *learning* – that is, changing perceptions over time regarding the structure of the system.

During recent years, analysis of learning behaviour has developed into an important and influential ingredient in reasoning about expectations, in two different respects. First, it is a much-noted fact that in most monetary macro-models there is a multiplicity (that is, two or more) of rational expectations (RE) solutions that are dynamically stable, that is, non-explosive. (Dynamically unstable solutions can usually be ruled out by recognition of a transversality condition that is relevant in the explicit or implicit optimization problem solved by the model's agents.) A widely held point of view is that in cases of indeterminacy – that is, two or more stable RE solutions – there is a presumption of substantial non-optimality for that reason alone, so that (for example) a monetary policy rule that permits such indeterminacy should be strenuously avoided (see, for example, Benhabib and Farmer 1999; Woodford 2003.) An alternative possibility, expressed most explicitly by McCallum (2003), is that in such cases only a single stable RE solution may be economically relevant. From that perspective there is evidently good reason to believe that a necessary (not sufficient!) condition, for a particular RE solution to be plausible, is that it be *learnable* by some process that enables individual agents in an economy to obtain empirical information about the parameters that govern the behaviour of the economy. (It seems implausible that individuals could obtain such information by processes that do not involve inference from data generated by the economy.) Extensive study of such processes has been one feature of a large body of work summarized in the influential treatise of Evans and Honkapohja (2001). The leading contender for a learning process for this first purpose is recursive least squares learning, with the issue being whether such a learning process converges to a RE solution as time passes and an unlimited quantity of data is accumulated. There exist alternative learning algorithms, of course, but the one in question is in several respects specified so as to be highly conducive to learnability, so that if a RE solution is not learnable by this procedure then it should not be regarded as a plausible candidate for an economically relevant equilibrium. This point of view may, in some cases, eliminate concerns regarding solution multiplicity. In this first type of learning analysis, the concept of E-stability – due to DeCanio (1979), Evans (1986), Marcet and Sargent (1989), and Evans and Honkapohja (1992) – is used extensively, as it provides a convenient technique for determining the learnability status of particular RE solutions. A notable application to monetary policy rules is Bullard and Mitra (2002).

A second, quite different, and more ambitious application of learning algorithms is to represent distinct hypotheses, as alternatives to RE, concerning the formation of expectations. This line of work has been pursued extensively by Evans and Honkapohja (2001) and numerous other researchers including Sargent (1993) and Orphanides and Williams (2005). An attractive feature of this approach is that it views expectational behaviour as departing somewhat from full, strict expectational rationality, but nevertheless retaining much of the intellectual discipline imposed by the RE requirement that behaviour regarding expectation formation be governed by the same optimizing benchmark that characterizes neoclassical economics more generally. From the substantive perspective, the use of learning processes (such as constant-gain modifications of least-squares learning that have the effect of down-weighting older data) in place of RE in calibrated macroeconomic models often gives rise to additional serial correlation in endogenous variables,

thereby generating model properties that are viewed by many analysts as being more nearly consistent with actual macro time-series data.

Historical Considerations

The foregoing discussion is somewhat historical in nature yet includes no references to literature predating the Second World War. Is there some explanation for this absence? The most celebrated discussion of inflationary expectations in the 'pre-war' literature is, probably, that of Irving Fisher. In *Appreciation and Interest* (1896), Fisher emphasizes the real versus nominal interest rate distinction that is often associated with his name, and in *The Theory of Interest* (1930) he estimates a distributed-lag regression relating interest to current and past inflation rates (interpreting the long lags as due to 'delayed adjustment'). In addition, several other economists (such as Marshall 1890) devoted some attention to the effects of expected inflation, the contribution of Henry Thornton (1802) being perhaps the most prescient (on this topic, see Humphrey 1983). All in all, however, it seems that the subject attracted little attention in the pre-war literature. Even in Knut Wicksell's famous analysis of the 'cumulative process' of inflation, there is only brief passing mention (1898, pp. 96, 148) of the possibility that the inflation will be anticipated. Discussion of the effects of sustained inflationary expectations on capital formation seems to be entirely absent. This neglect may perhaps be satisfactorily explained by first noting that it is sustained inflation that is relevant and then recalling that during this earlier era the world's major economies normally adhered to some commodity- money standard, thereby sharply reducing the scope for substantial inflation to arise or to be sustained.

See Also

- ▶ Adaptive Expectations
- ▶ Expectations
- ▶ Inflation
- ▶ Inflation Dynamics
- ▶ Learning in Macroeconomics
- ▶ Lucas Critique
- ▶ New Classical Macroeconomics
- ▶ Phillips Curve
- ▶ Rational Expectations

Bibliography

Bailey, M.J. 1956. The welfare cost of inflationary finance. *Journal of Political Economy* 64 (2): 93–110.

Barro, R.J., and S. Fischer. 1976. Recent developments in monetary theory. *Journal of Monetary Economics* 2: 133–167.

Benhabib, J., and R.E.A. Farmer. 1999. Indeterminacy and sunspots in macroeconomics. In *Handbook of macroeconomics*, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Bernanke, B.S., T. Laubach, F.S. Mishkin, and A.S. Posen. 1999. *Inflation targeting: Lessons from the international experience*. Princeton: Princeton University Press.

Bullard, J., and K. Mitra. 2002. Learning monetary policy rules. *Journal of Monetary Economics* 49: 1105–1129.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Calvo, G.A. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Clarida, R., J. Gali, and M. Gertler. 1999. The science of monetary policy: A new Keynesian perspective. *Journal of Economic Literature* 37: 1661–1707.

DeCanio, S. 1979. Rational expectations and learning from experience. *Quarterly Journal of Economics* 94: 47–57.

Eggertsson, G.B., and M. Woodford. 2003. The zero bound on interest rates and optimal monetary policy. *Brookings Papers on Economic Activity* 2003 (1): 139–233.

Evans, G.W. 1986. Selection criteria for models with non-uniqueness. *Journal of Monetary Economics* 18: 147–157.

Evans, G.W., and S. Honkapohja. 1992. On the robustness of bubbles in linear RE models. *International Economic Review* 33: 1–14.

Evans, G.W., and S. Honkapohja. 2001. *Learning and expectations in macroeconomics*. Princeton: Princeton University Press.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85: 191–205.

Fisher, I. 1896. *Appreciation and interest*. New York: American Economic Association.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Fisher, M.E., and J.J. Seater. 1993. Long-run neutrality and superneutrality in an ARIMA framework. *American Economic Review* 83: 402–415.

Friedman, M. 1960. *A program for monetary stability*. New York: Fordham University Press.
Friedman, M. 1966. Comments. In *Guidelines, informal controls, and the market place*, ed. G.P. Shultz and R.Z. Aliber. Chicago: University of Chicago Press.
Friedman, M. 1968. The role of monetary policy. *American Economic Review Papers and Proceedings* 58: 1–18.
Friedman, M. 1969. *The optimum quantity of money and other essays*. Chicago: Aldine.
Goodfriend, M., and R.G. King. 1997. The new neoclassical synthesis and the role of monetary policy. In *NBER macroeconomics annual 1997*, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.
Goodfriend, M., and R.G. King. 2005. The incredible Volcker disinflation. *Journal of Monetary Economics* 52: 981–1015.
Gordon, R.J. 1976. Recent developments in the theory of inflation and unemployment. *Journal of Monetary Economics* 2: 185–219.
Humphrey, T.M. 1983. The early history of the real/nominal interest rate relationship. *Federal Reserve Bank of Richmond Economic Review* 69 (3): 2–10.
King, R.G., and M.W. Watson 1992. *Testing long-run neutrality*. Working paper no. 4156. Cambridge, MA: NBER.
Lucas, R.E. Jr. 1972a. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
Lucas, R.E. Jr. 1972b. Econometric testing of the natural rate hypothesis. In *The econometrics of price determination conference*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.
Lucas, R.E. Jr. 1973. Some international evidence on output inflation tradeoffs. *American Economic Review* 63: 326–334.
Mankiw, N.G., and R. Reis. 2002. Sticky information versus sticky prices: A proposal to replace the new Keynesian Phillips curve. *Quarterly Journal of Economics* 117: 1295–1328.
Marcet, A., and T.J. Sargent. 1989. Convergence of least-squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory* 48: 337–368.
Marshall, A. 1890. *Principles of economics*. London: Macmillan.
McCallum, B.T. 1976. Rational expectations and the natural rate hypothesis: Some consistent estimates. *Econometrica* 44: 43–52.
McCallum, B.T. 1984. Low-frequency estimates of long-run relationships in macroeconomics. *Journal of Monetary Economics* 14: 3–14.
McCallum, B.T. 1990. Inflation: Theory and evidence. In *Handbook of monetary economics*, ed. B.M. Friedman and F.H. Hahn. Amsterdam: North-Holland.
McCallum, B.T. 2003. Multiple-solution indeterminacies in monetary policy analysis. *Journal of Monetary Economics* 50: 1153–1175.
Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.
Orphanides, A., and J.C. Williams. 2005. Imperfect knowledge, inflation expectations, and monetary policy. In *The inflation targeting debate*, ed. B.S. Bernanke and M. Woodford. Chicago: University of Chicago Press.
Phelps, E.S. 1967. Phillips curves, expectations of inflation, and optimal unemployment over time. *Economica* 34: 254–281.
Rotemberg, J.J., and M. Woodford. 1997. An optimization-based econometric framework for the evaluation of monetary policy. In *NBER macroeconomics annual 1997*, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.
Sargent, T.J. 1971. A note on the accelerationist controversy. *Journal of Money, Credit and Banking* 3: 721–725.
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.
Sargent, T.J. 1993. *Bounded rationality in macroeconomics*. Oxford: Oxford University Press.
Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review* 57: 534–544.
Solow, R.M. 1969. *Price expectations and the behavior of the price level*. Manchester: Manchester University Press.
Svensson, L.E.O., and M. Woodford. 2005. Implementing optimal policy through inflation-forecast targeting. In *The inflation-targeting debate*, ed. B.S. Bernanke and M. Woodford. Chicago: University of Chicago Press.
Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.
Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of Great Britain*, ed. F.A. von Hayek. Fairfield: Kelley, 1978.
Whiteman, C.H. 1984. Lucas on the quantity theory: Hypothesis testing without theory. *American Economic Review* 74: 742–749.
Wicksell, K. 1898. *Interest and prices*. Trans. R.F. Kahn. London: Macmillan, 1936.
Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.

Inflation Measurement

David E. Lebow and Jeremy B. Rudd

Abstract

Inflation measurement is the process whereby changes in the prices of individual goods and services are combined to yield a measure of general price change. This article discusses the conceptual framework for thinking about

inflation measurement and considers practical issues associated with determining an inflation measure's scope; with measuring individual prices; and with combining these individual prices into a measure of aggregate inflation. We also discuss the concept of 'core inflation' and summarize the implications of inflation measurement for economic theory and policy.

Keywords

Aggregate price level definition; Asset price measurement; Cobb–Douglas functions; Conditional cost-of-living index; Constant elasticity of substitution; Consumer price index; Core inflation; Cost-of-goods index; Cost-of-living index; Dynamic factor models; 'Exclusion' measures of inflation; Fisher index; GDP price index; Headline inflation; Hedonic price functions; Index numbers; Indexation for inflation; Inflation measurement; Laspeyres index; Lifetime utility; Limited-influence measures of inflation; Matched model price index; Measurement error; 'Neo-Edgeworthian' index; Paasche index; Quality-adjustment problem; Reduced-form Phillips curve; Rental equivalence; Shadow prices; 'Stochastic' approach to inflation measurement; Törnqvist index; Underlying inflation

JEL Classifications

C43; E31

Inflation measurement is the process whereby changes in the prices of individual goods and services are combined to yield a measure of general price change. In formal terms, we may specify the time-*t* rate of aggregate inflation *Pt* as

$$P_t = F\left(p_t^1, p_t^2, \ldots p_t^I\right), \tag{1}$$

where $F(.)$ is a function that aggregates a set of I individual time-t price changes p_t^i.Writing the problem in this manner highlights three basic issues associated with inflation measurement. First, we must decide what collection of price changes we wish to include (or, more generally, what should be the measure's *scope*); second, we must ensure that the *individual* price changes are correctly measured; and finally, we must choose a method for combining those changes into a measure of aggregate inflation.

While the problem of inflation measurement can be broadly described in these terms, dealing with the numerous complications that emerge in practice requires some explicit conceptual framework. Probably the simplest way to construct a measure of overall inflation involves defining the aggregate price level in terms of the cost of a fixed basket of goods and services. Such a measure – sometimes labeled a *cost-of-goods index* (COGI) – has several practical advantages; in particular, for a broad enough basket of goods, the change in a COGI comes very close to what most people intuitively mean by an inflation rate, and a COGI-type measure can easily be defined for any sub-component of expenditure or production (such as consumption, investment, or the output of intermediate goods). However, this simple measure of inflation faces an important practical difficulty. In a dynamic economy, the composition and nature of output will evolve as existing goods are consumed or produced in different quantities, as the characteristics of existing products change, or as entirely new goods are introduced; these changes make the COGI's fixed bundle of goods become less representative over time. The COGI approach provides no guidance as to how to address this problem, suggesting that a more comprehensive conceptual framework is needed.

If we confine our attention to consumption prices, then a natural guiding principle is provided by the concept of a *cost-of-living index* (COLI), which measures the expenditure needed for an optimizing consumer to maintain a specified level of utility as prices change. The strength of the COLI framework derives from its grounding in the theory of consumer behaviour, which can provide clear-cut suggestions (at least in principle) as to how to deal with such problems as changes in expenditure patterns or the introduction of new goods. That said, this feature of the COLI approach can also be a weakness to the extent that consumer theory provides an incorrect characterization of actual behaviour (NRC 2002,

pp. 53–8) or is insufficiently well developed to handle a particular practical situation. In addition, because the COLI concept pertains only to consumption, it provides little or no guidance about the construction of broader measures of inflation that include prices for other components of output (these might be of interest, for instance, to a monetary policymaker). The COLI framework therefore provides a natural guide to the construction of a consumer price index (CPI), which attempts to measure the prices of goods and services consumed by households; it will not, however, be able to inform the construction of a price index for overall GDP, which is defined to include the prices of *all* domestically produced final output – whether purchased by consumers, businesses, governments, or the rest of the world. (While a literature does exist on the measurement of price change from a *producer* perspective – see Diewert 1983, for an overview – it has generally received much less attention than the corresponding consumer-based approach.)

Despite these potential shortcomings, a COLI-based approach is commonly employed as a framework for informing inflation measurement – in the United States, for example, the CPI uses the COLI concept both as its explicit measurement objective and as a reference for making practical decisions about index construction (U.S. Bureau of Labor Statistics 2005). In much of what follows, therefore, we follow common practice in using the concept of a cost-of-living index to guide our discussion of the three basic issues – scope, individual price measurement, and aggregation – that are associated with the measurement of inflation. In addition, we discuss the concept of *core inflation*, which can be motivated and interpreted in terms of an alternative approach to inflation measurement. Finally, we conclude by considering the implications of these measurement issues for economic research and policy.

What Items Should Be Included in an Inflation Measure?

The scope or domain of a cost-of-goods index – whether it is defined for consumption goods or more broadly – is defined to include all items that are purchased and sold in market transactions, and, hence, that have well-defined prices. (In reality, of course, any inflation measure will include only a subset of goods consumed or produced in the economy, so sampling in order to provide a representative characterization of aggregate price change represents an important practical concern.) By contrast, the scope of a cost-of-living index is much broader than that of a corresponding COGI for consumption goods in as much as a COLI needs to account for *anything* that affects utility, including changes over time in 'background' or 'environmental' factors such as weather, pollution, crime, or the provision of public goods.

For a COLI-based measure of consumption price inflation, therefore, the relevant set of price changes $p_t^1, p_t^2, \ldots p_t^I$ should in principle include changes in both market prices and the 'shadow prices' of environmental factors (with the latter defined in the sense of Pollak 1989). In practice, however, it is almost impossible to correctly measure the effect on utility of these sorts of changes (even if we could do so, inclusion of such factors strays beyond what most people understand by the term 'inflation'). These considerations lead to the concept of a 'conditional' COLI, which (to follow Pollak 1989, again) is defined as the smallest change in expenditure that is required in order to maintain a reference utility level following a change in prices, *with the state of the environment fixed*. Although intuitive, the concept of a conditional COLI has its own conceptual difficulties. In particular, since preferences over market goods will likely depend on the environment (for example, demand for medical care depends on the incidence of disease), the rate of inflation implied by a conditional COLI will depend on the particular state of the environment that we condition on.

While the concept of a conditional COLI provides useful guidance regarding the relevant domain of a measure of consumption price inflation, it cannot unambiguously solve all questions about scope. For example, many households receive an implicit flow of services from owner-occupied housing. On the assumption that the 'price' of these services could be measured, it is

unclear whether they should enter a COLI given that they are not generated by a market transaction or explicit expenditure (and are not closely related to a conventional notion of a price); of course, the initial home purchase does meet these criteria. A similar problem extends to a number of other goods that are consumed by households but not directly purchased by them (one example is banking services furnished without explicit charge, which are included in most national accounts' definitions of consumption). And, again, once one moves outside of the realm of private consumption, the conditional COLI framework provides no practical guidance regarding the construction of inflation measures for other components of production or spending (such as investment) or for broader measures of inflation (such as the GDP price index).

Finally, a particularly difficult and controversial issue concerns the proper role of asset prices in inflation measures. If we extend the theory of a cost-of-living index to an intertemporal or multiperiod context (see Pollak 1975), then expected changes in the price of future consumption streams can affect current inflation through their impact on *lifetime* utility. We can therefore consider a cost-of-living index that is defined to include current and future prices of consumption goods; furthermore, to the extent that information about future consumption prices is contained in current asset prices, an argument can be made for including these prices in a COLI-based inflation measure (Alchian and Klein 1973). In practice, however, the volatility of asset prices – as well as the related fact that observed movements in asset prices can stem from sources unrelated to expected future consumption-price changes – typically precludes their inclusion in conventional inflation measures. (The current purchase prices of durable goods, which *are* often included, provide a partial exception.)

How Should Individual Price Changes Be Measured?

A number of practical problems complicate the measurement of individual price changes. First, in a modern economy the characteristics of existing goods can change over time; likewise, new goods and services will constantly be entering – and old goods leaving – production and consumption. Left unaddressed, these problems will render it impossible to track the price changes for an identical set of goods and will cause the set of goods being priced to become increasingly less representative of actual consumption and production. This will obviously affect COGI-based measures of inflation, and it will also affect COLI-based measures to the extent that changes in the characteristics or variety of available goods have an effect on the utility that is realized from their consumption.

Several techniques exist for dealing with non-trivial changes in the characteristics (loosely speaking, the 'quality') of existing products; all of these involve some procedure for dividing the observed price change into a component that reflects changes in the good's characteristics and a component that reflects 'pure' price change, where only this latter component is appropriate for inclusion in an inflation measure. (Moulton and Moses 1997 and NRC 2002, provide a detailed description and assessment of these various methods of quality adjustment in the context of the US CPI; see also ILO et al. 2004.) For example, when the original and modified products exist in the same period, any difference in their prices can be attributed to differences in the goods' characteristics. Alternatively, in the more common case where a good exists in one form in period t and in another in period $t + 1$, the 'pure' price change over the intervening period can be imputed from the observed average price change for a similar group of goods. (A 'matched model' index, which only includes price changes for goods that remain in the sample without change – and so implicitly assigns that average price change to other items – is a common example.) Finally, additional information may be brought to bear on the problem: under certain assumptions, for example, data on the cost to manufacturers of modifying the characteristics of a product can be used to compute the effect of these modifications on the good's price.

When detailed information about a product's characteristics is available, so-called 'hedonic' methods may be used. The hedonic approach relates the observed price of a good to its characteristics; any change in characteristics can then be explicitly controlled for and removed from the good's total price change. Specifically, when the individual effects of a good's characteristics on its price are stable over time, a measure of pure price change can be obtained by permitting the level of the price-characteristics relation to shift in each period. In the more realistic case where different hedonic functions exist for each period, a measure of pure price change between periods t and $t + 1$ can be defined as $h^{t+1}(z_t)/h^t(z_t)$ *or* $h^{t+1}(z_{t+1})/h^t(z_{t+1})$, where $h^i(z)$ denotes the hedonic function in period i relating the good's price to its set of characteristics z. (Here, the first expression yields a 'Laspeyres-like' price measure as the hedonic function is evaluated with the set of characteristics from the variety that is purchased in the base period; similarly, the second expression yields a 'Paasche-like' measure.)

An important advantage of the hedonic approach to dealing with quality change is that it can be explicitly grounded in cost-of-living theory. Under relatively weak conditions, $h^{t+1}(z^t) - h^t(z_t)$ provides an upper bound for the compensating variation associated with a given price change; likewise, $h^{t+1}(z_{t+1}) - h^t(z_{t+1})$ gives a lower bound for the equivalent variation (NRC 2002, pp. 153–4). It is unknown, however, whether these bounds are particularly tight. In addition, statistical agencies typically find real-time production of measures like these too difficult, and instead produce quality-adjusted price changes by scaling the observed price change for a good by the ratio $h^{t-j}(z_t)/h^{t-j}(z_{t+1})$, where the $t-j$ superscript makes apparent the dependence of the estimated hedonic function on an earlier period's data. Such a procedure cannot, in general, be justified in terms of a COLI-based approach (Pakes 2002).

The 'new goods' problem can be thought of as a more difficult variant of the quality-adjustment problem in which the new good contains features or characteristics that have never existed before (in a sense, the dimension of the 'characteristics space' has increased): examples include the introduction of the video cassette recorder or cellular telephone. In this case, one needs a method for imputing the price of a newly introduced good in the period prior to its first appearance in the economy; as was suggested by Hicks (1940, pp. 114–15), one logical imputation involves setting this pre-introduction price equal to the price at which the demand for the good is just equal to zero. While such an approach can be explicitly motivated in terms of a COLI-based framework, its implementation requires a degree of information about consumer preferences that is unlikely to be realized in practice (see Hausman 1997, for a representative example). It is therefore common for statistical agencies to attempt to mitigate the new goods problem through the more rapid addition of new items into the set of price changes being tracked over time; while intuitive, this approach may not always ameliorate the effects of new-goods introduction (Pakes 2002).

Another problem that arises in measuring individual price changes relates to the fact that even identical goods can sell for different prices across different sellers. These differentials could reflect true price differences – a particular outlet might simply be able to charge a lower price – but they could also reflect characteristics of the outlet itself, such as customer service or convenience. In the latter case, two otherwise identical goods should be treated as different products if they are sold at different outlets; similarly, when the outlet used to price a particular good changes, some adjustment – akin to the sorts of quality adjustments discussed above – must be made to the good's price.

One final issue relating to the measurement of individual price changes is that a good's purchase price need not be related to its effect on current-period utility if it provides consumption services in more than one period (as is the case for a durable good) or if it can be stored for later consumption. For a durable good, the conceptually relevant measure of the change in the good's price in a given period is the change in its user cost. In practice, the user cost turns out to be difficult to estimate and often implies erratic price movements. In the presence of a well-functioning rental

market, the cost of hiring a good can serve as a proxy for its user cost; this 'rental equivalence' procedure is used in the US CPI for owner-occupied housing. However, the absence of rental markets for most durable goods limits the usefulness of this technique, and in practice the purchase prices of many durable goods are directly included in most inflation measures.

How Should the Individual Price Changes Be Combined?

The combination of individual price changes into an aggregate measure of inflation falls into the domain of the theory of *index numbers*, a full discussion of which is beyond the scope of this survey. We therefore focus on some of the practical issues that arise in choosing and implementing an aggregation formula.

A natural way to construct a cost-of-goods index involves weighting the individual price changes for the components of the fixed market basket by their shares in overall expenditures. When the initial period of the index is the same as the period used to specify the expenditure weights, the resulting measure corresponds to a Laspeyres index. As is well known, however, a Laspeyres index overstates changes in the cost of living when consumer substitution occurs in response to changes in relative prices; hence, alternative formulas that do capture substitution behaviour can provide a more accurate approximation to a COLI. Examples include the Törnqvist and Fisher ideal indexes (both members of the 'superlative' class of index numbers defined by Diewert 1976), which employ aggregation weights derived from quantities purchased in both the initial and final periods of the comparison. Although the theory is not as well developed as that for consumer expenditures, similar justifications for commonly employed superlative aggregation formulas may exist for broader measures of output prices as well (for example, see Diewert 1983, for a production-based interpretation of a Törnqvist index).

In addition, statistical agencies often make use of 'chaining' (Fisher 1911, ch. 10; Forsyth and Fowler 1981) when constructing long time series of inflation rates; with this procedure, the price changes implied by a sequence of indexes defined over various sub-periods are 'chained' or cumulated together. In the COGI context, chaining carries an intuitive or pragmatic appeal in as much as it ensures that the basket being priced will remain reasonably representative of actual consumption patterns over time. However, chaining by itself cannot correctly capture consumer substitution. (Feenstra and Shapiro 2003, and Szulc 1983, consider other problems that can arise with chained indexes.)

In many circumstances, price indexes must be constructed in the absence of timely data on expenditures. A superlative aggregation formula cannot be used in real time in these cases (indeed, the fact that the Laspeyres index requires only expenditures from an earlier base period accounts for much of its appeal). A compromise procedure, which requires only base-period expenditure data, involves using a weighted constant elasticity of substitution (CES) aggregator (this includes the weighted geometric mean – which measures the cost of living when utility takes a Cobb–Douglas form – as a special case). Based on historical evidence, one could form a judgement about the likely degree of substitutability across items and then use an appropriately calibrated CES formula (Shapiro and Wilcox 1997). Such a procedure is now employed by the US CPI to aggregate *individual* prices (that is, prices *within* item-area strata), with a geometric means formula used for the majority of cases and a Laspeyres formula reserved for strata where substitution is deemed unlikely a priori.

Accurately capturing substitution behaviour is not the only relevant issue for choosing an aggregation formula. Statistical agencies typically measure a sample of prices (where the number of price quotes for a given sub-index may be quite small), and commonly used formulas can differ in their susceptibility to small-sample biases. Indeed, Bradley (2001) has argued that

small-sample bias in Laspeyres indexes – not a failure to capture substitution across categories of goods – accounts for most of the observed difference between the published (Laspeyres) version of the US CPI and a superlative (Törnqvist) variant.

At least two other issues arise in choosing how to combine individual price changes into a measure of overall inflation. First, the weights selected for use in aggregation can reflect explicit or implicit judgements as to which agents are to be represented in the index. By employing aggregate expenditure weights, the typical consumer price measure in effect gives a larger weight to the inflation rates faced by richer households – a so-called 'plutocratic' weighting scheme. (Alternatively, we could compute the simple average of each household-specific inflation rate; this 'democratic' weighting scheme might be more representative of a 'typical' household's experience.) For some purposes, one might also explicitly choose to measure the inflation rate faced by a particular segment of the population, such as wage earners, the poor, or the elderly.

Second, correct measurement of the quantities used in aggregation is critical. To the extent that these are subject to measurement error (as might occur if they are estimated from survey data), and to the extent that mismeasured weights are systematically associated with items that display above-or below-average price changes, the resulting aggregate inflation rate will be mismeasured. (Lebow and Rudd 2003, present evidence of this in the US CPI.)

The Concept of Core Inflation

Core inflation was originally defined as 'the trend rate of increase' of either 'the price of aggregate supply' or 'the cost of the factors of production' (Eckstein 1981). More commonly, however, core inflation is understood in a statistical sense as corresponding either to 'underlying inflation' (the portion of overall inflation that is free from transitory influences) or to a measure of the common trend in all prices. In line with its various definitions, core inflation can be measured in a variety of ways.

The most prevalent core inflation measures are 'exclusion' measures that omit certain items, such as food and energy, from the calculation of overall inflation. The popularity of excluding food and energy derives in part from the experience of the 1970s and early 1980s, which saw sizable supply-driven price hikes for these items. Many prices other than food and energy may move erratically as well, however (indeed, some countries publish exclusion-based measures of core consumer price inflation that omit housing, the effects of changes in indirect taxes, or other items). Thus, a variant on the exclusion approach involves adjusting the weight of items in inverse proportion to their variability (sometimes termed a 'neo-Edgeworthian' index), so that items with erratic prices are downweighted rather than omitted entirely.

A second category of core inflation measures includes limited-influence measures such as medians or trimmed means (Bryan and Cecchetti 1994). These measures exclude a certain proportion of the largest and smallest price changes each period (in the extreme case of the median, all items but one are excluded each period). In contrast to standard exclusion measures, however, the omitted items will vary period by period. Limited-influence measures sometimes do well in statistical exercises aimed at finding measures that are well correlated with long moving averages of headline inflation, or measures that can serve as good univariate predictors of headline inflation. However, for these limited-influence measures to capture underlying inflation well, true relative price changes must be smaller than transitory fluctuations (which will not always be the case). In addition, construction of these measures is often sensitive to the degree of disaggregation employed and to the length of time over which the individual price changes are measured.

A third set of approaches uses econometric techniques to estimate core inflation (variously

defined). For example, in an econometric reduced-form Phillips curve (as was employed in Eckstein's original study), lagged inflation terms can proxy for the persistent component of inflation once one controls for supply shocks and aggregate demand (the univariate analogue would involve taking simple or weighted averages of past inflation as the core inflation measure). Another approach is to use a dynamic factor model to extract a common component or 'signal' from a set of disaggregated inflation rates (Bryan and Cecchetti 1993). Other econometric approaches have been proposed as well (often invoking economic theory to provide their rationale) – for example, core inflation may be defined as the component of inflation that is uncorrelated with long-run economic activity (Quah and Vahey 1995), or best correlated with money growth. Of course, these theory-based underpinnings might be controversial; more generally, econometric approaches might be difficult to understand or communicate.

The neo-Edgeworthian, limited-influence, and dynamic factor approaches to measuring core inflation exemplify an alternative 'statistical' or 'stochastic' approach to inflation measurement that has garnered increased interest in recent years (Wynne 1997). Wynne contends that the economic basis for these inflation concepts is 'some concept of "monetary" inflation that...is not necessarily the same thing as changes in the cost of living'. If so, these alternative approaches will in principle imply different decisions about scope and aggregation relative to those implied by a COLI-based framework. In particular, to the extent that these measures seek to capture the portion of aggregate price movement that is attributable to changes in the supply of money, their relevant scope could be the price of *any* transaction that involves an exchange of money (including prices for financial assets and the purchase prices – not the user costs – for durable goods). In addition, the aggregation weights employed by these stochastic approaches are typically informed by purely statistical considerations, and so need not bear any resemblance to the weights implied by cost-of-living theory.

Implications for Research and Policy

Inflation measurement matters for at least three reasons. First, and most obviously, economic decisions often depend directly – even automatically – on published inflation measures. In the public sphere, many government programmes are indexed to inflation measures such as changes in a consumer price index: in the United States, for example, Social Security benefits, income tax schedules, and coupon payments on inflation-indexed government debt are all directly tied to changes in the CPI. Private contracts, including wage arrangements, are also indexed to changes in the CPI (although such indexation provisions are less common today than they were when inflation was higher and more uncertain).

The use of inflation measures in indexation arrangements, in principle, should help inform the details of inflation measurement. If indexation of a payment is intended to maintain its real purchasing power for a recipient, then this goal is best served by using an inflation measure tailored to that recipient. Thus, indexation of pension payments would utilize an inflation measure that reflects the consumption patterns of pensioners; income-support payments would use a measure reflecting the consumption of the poor; and so on. Such specialized price indexes can differ from an aggregate price index in both the choice of priced items and in the weights assigned to them.

Inflation measurement is also important because inflation affects economic welfare and therefore serves as a goal of public policy in its own right – in particular, a central objective of monetary policymakers is the maintenance of low and stable inflation. Problems measuring the average level of inflation will therefore affect a central bank's choice of inflation target (whether explicit or implicit). For example, many argue that the Federal Reserve should seek to stabilize *measured* inflation at some level higher than zero, in part because the US CPI tends to overstate changes in the cost of living (for example, Bernanke et al. 1999). More problematically, if measurement errors in inflation vary over time in unknown ways, central banks could respond inappropriately to movements in observed inflation rates.

Finally, because real quantities are typically estimated by deflating nominal values with a price index, inflation measurement directly affects the construction of other economic statistics (including real GDP and productivity). Thus, our ability to correctly assess the effects of technological progress, the sources of economic growth and changes in living standards over time hinges in an obvious way on the accurate measurement of individual and aggregate price movements. Furthermore, if the extent of measurement error in inflation varies over time and across items or places, then growth comparisons could be affected; examples include measuring changes in living standards over long periods (Gordon 2005) and comparing growth and productivity performance in the United States and Europe (Ahmad et al. 2003).

See Also

- ▶ Hedonic Prices
- ▶ Index Numbers

Bibliography

Ahmad, N., F. Laquiller, P. Marianna, D. Pilat, P. Schreyer, and A. Wölfl. 2003. Comparing labour productivity growth in the OECD area: The role of measurement. OECD Science, Technology and Industry Statistics working papers, 2003/14. OECD Publishing. doi: 10.1787/126534183836.

Alchian, A., and B. Klein. 1973. On a correct measure of inflation. *Journal of Money, Credit, and Banking* 5: 173–191.

Bernanke, B., T. Laubach, F. Mishkin, and A. Posen. 1999. *Inflation targeting: Lessons from the international experience*. Princeton: Princeton University Press.

Bradley, R. 2001. Finite sample effects in the estimation of substitution bias in the consumer price index. *Journal of Official Statistics* 17: 369–390.

Bryan, M., and S. Cecchetti. 1993. The consumer price index as a measure of inflation. *Federal Reserve Bank of Cleveland Economic Review* 29: 15–24.

Bryan, M., and S. Cecchetti. 1994. Measuring core inflation. In *Monetary policy*, ed. N. Mankiw. Chicago: University of Chicago Press.

Diewert, W. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4: 115–145.

Diewert, W. 1983. The theory of the output price index and the measurement of real output change. In *Price level measurement*, ed. W. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Eckstein, O. 1981. *Core inflation*. Englewood Cliffs: Prentice-Hall.

Feenstra, R., and M.D. Shapiro. 2003. High-frequency substitution and the measurement of price indexes. In *Scanner data and price indexes*, ed. M. Shapiro and R. Feenstra. Chicago: University of Chicago Press.

Fisher, I. 1911. *The purchasing power of money*. New York: Macmillan.

Forsyth, F., and R. Fowler. 1981. The theory and practice of chain price index numbers. *Journal of the Royal Statistical Society (Series A)* 144: 224–246.

Gordon, R. 2005. Apparel prices 1914–93 and the Hulten/Brueghel paradox. Working paper no. 11548. Cambridge, MA: NBER.

Hausman, J. 1997. Valuation of new goods under perfect and imperfect competition. In *The economics of new goods*, ed. T. Bresnahan and R. Gordon. Chicago: University of Chicago Press.

Hicks, J. 1940. The valuation of the social income. *Economica* 7: 105–124.

ILO/IMF/OECD/UNECE/Eurostat/World Bank. 2004. *Consumer price index manual: Theory and practice*. Geneva: International Labour Office.

Lebow, D., and J. Rudd. 2003. Measurement error in the consumer price index: Where do we stand? *Journal of Economic Literature* 41: 159–201.

Moulton, B., and K. Moses. 1997. Addressing the quality-change issue in the consumer price index. *Brookings Papers on Economic Activity* 1997(1): 305–366.

NRC (National Research Council). 2002. *At what price? Conceptualizing and measuring cost-of-living and price indexes*. Washington, DC: National Academy Press.

Pakes, A. 2002. A reconsideration of hedonic price indices with an application to PCs. Working paper no. 8715. Cambridge, MA: NBER.

Pollak, R. 1975. The intertemporal cost of living index. *Annals of Economic and Social Measurement* 4: 179–195.

Pollak, R. 1989. The treatment of the environment in the cost-of-living index. In *The theory of the cost of living index*, ed. R. Pollak. New York: Oxford University Press.

Quah, D., and S. Vahey. 1995. Measuring core inflation. *Economic Journal* 105: 1130–1144.

Shapiro, M., and D. Wilcox. 1997. Alternative strategies for aggregating prices in the CPI. *Federal Reserve Bank of St. Louis Review* 79(3): 113–125.

Szulc, B. 1983. Linking price index numbers. In *Price level measurement*, ed. W. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

U.S. Bureau of Labor Statistics. 2005. The consumer price index. In *BLS handbook of methods*. Washington, DC: GPO.

Wynne, M. 1997. Commentary. *Federal Reserve Bank of St. Louis Review* 79(3): 161–167.

Inflation Targeting

Lars E. O. Svensson

Abstract

Inflation targeting is a monetary-policy strategy that was introduced in New Zealand in 1990, has been very successful in terms of stabilizing both inflation and the real economy, and as of 2007 had been adopted by more than 20 industrialized and nonindustrialized countries. It is characterized by an announced numerical inflation target, an implementation of monetary policy that gives a major role to an inflation forecast and has been called 'inflation-forecast targeting', and a high degree of transparency and accountability.

Keywords

Accountability; Central banking; Central banks; Consumer price index; Core consumer price index; Credibility; Fixed exchange rates; Flexible exchange rates; Flexible inflation targeting; Forecast targeting; Forecasting; Inflation expectations; Inflation targeting; Liquidity trap; Monetary policy; Monetary policy instrument; Monetary policy rules; Monetary transmission mechanism; Money supply; Money-growth targeting; Optimal instrument-rate paths; Output gap; Potential output; Price stability; Price-level targeting; Stabilization policy; Transparency; Unit root

JEL Classifications

E3

Inflation targeting is a monetary-policy strategy that was introduced in New Zealand in 1990, has been very successful, and as of 2007 had been adopted by more than 20 industrialized and non-industrialized countries. It is characterized by (a) an announced numerical inflation target, (b) an implementation of monetary policy that gives a major role to an inflation forecast and has been called 'inflation-forecast targeting', and (c) a high degree of transparency and accountability.

The *numerical inflation target* is typically around two per cent at an annual rate for the Consumer Price Index (CPI) or a core CPI, in the form of a range, such as one to three per cent in New Zealand; or a point target with a range, such as a two per cent point target with a range/tolerance interval of plus/minus one percentage points in Canada and Sweden; or a point target without any explicit range, such as two per cent in the UK and 2.5 per cent in Norway. The difference between these forms does not seem to matter in practice: a central bank with a target range seems to aim for the middle of the range, and the edges of the range are normally interpreted as 'soft edges' in the sense that they do not trigger discrete policy changes, and being just outside the range is not considered much different from being just inside.

In practice, inflation targeting is never 'strict' inflation targeting but always 'flexible' inflation targeting, in the sense that all inflation-targeting central banks ('central bank' is used as the generic name for monetary authority) not only aim at stabilizing inflation around the inflation target but also put some weight on stabilizing the real economy, for instance, implicitly or explicitly stabilizing a measure of resource utilization such as the output gap between actual output and 'potential' output. Thus, the 'target variables' of the central bank include not only inflation but other variables as well, such as the output gap. The objectives under flexible inflation targeting seem well approximated by a quadratic loss function consisting of the sum of the square of inflation deviations from target and a weight times the square of the output gap, and possibly also a weight times the square of instrument-rate changes (the last part corresponding to a preference for interest-rate smoothing). (The instrument rate is the short nominal interest rate that the central bank sets to implement monetary policy.) However, for new inflation-targeting regimes, where the establishment of 'credibility' is a priority, stabilizing the real economy probably has less

weight than when credibility has been established (more on credibility below).

Because there is a lag between monetary-policy actions (such an instrument-rate change) and its impact on the central bank's target variables, monetary policy is more effective if it is guided by forecasts. The implementation of inflation targeting therefore gives a main role to forecasts of inflation and other target variables. It can be described as *forecast targeting,* that is, setting the instrument rate (more precisely, deciding on an instrument-rate path) such that the forecasts of the target variables conditional on that instrument-rate path 'look good', where 'look good', for instance, means that the inflation forecast approaches the inflation target and the output-gap forecast approaches zero at an appropriate pace.

Inflation targeting is characterized by a high degree of *transparency.* Typically, an inflation-targeting central bank publishes a regular monetary-policy report which includes the bank's forecast of inflation and other variables, a summary of its analysis behind the forecasts, and the motivation for its policy decisions. Some inflation-targeting central banks also provide some information on, or even forecasts of, their likely future policy decisions.

This high degree of transparency is exceptional in view of the history of central banking. Traditionally, central-bank objectives, deliberations, and even policy decisions have been subject to considerable secrecy. It is difficult to find any reasons for that secrecy beyond central bankers' desire not to be subject to public scrutiny (including scrutiny and possible pressure from governments or legislative bodies). The current emphasis on transparency is based on the insight that monetary policy to a very large extent is 'management of expectations'. Monetary policy has an impact on the economy mostly through the private-sector expectations that current monetary-policy actions and announcements give rise to. The level of the instrument rate for the next few weeks matter very little to most economic agents. What matters is the expectations of future instrument settings, which expectations affect longer interest rates that do matter for economic decisions and activity.

Furthermore, private-sector expectations of inflation for the next one or two years affect current pricing decisions and inflation for the next few quarters. Therefore, the anchoring of private-sector inflation expectations on the inflation target is a crucial precondition for the stability of actual inflation. The proximity of private-sector inflation expectations to the inflation target is often referred to as the 'credibility' of the inflation-targeting regime. Inflation-targeting central banks sometimes appear to be obsessed by such credibility, there are good reasons for this obsession. If a central bank succeeds in achieving credibility, a good part of the battle to control inflation is already won. A high degree of transparency and high-quality and convincing monetary-policy reports are often considered essential to establishing and maintaining credibility. Furthermore, a high degree of credibility gives the central bank more freedom to be 'flexible' and also stabilize the real economy.

Whereas many central banks in the past seem to have actively avoided *accountability*, for instance by not having explicit objectives and by being very secretive, inflation targeting is normally associated with a high degree of accountability. A high degree of accountability is now considered generic to inflation targeting and an important component in strengthening the incentives faced by inflation-targeting central banks to achieve their objectives. The explicit objectives and the transparency of monetary-policy reporting contribute to increased public scrutiny of monetary policy. In several countries inflation-targeting central banks are subject to more explicit accountability. In New Zealand, the Governor of the Reserve Bank of New Zealand is subject to a Policy Target Agreement, an explicit agreement between the Governor and the government on the Governor's responsibilities. In the UK, the Chancellor of the Exchequer's remit to the Bank of England instructs the Bank to write a public letter explaining any deviation from the target larger than one percentage point and what actions the Bank is taking in response to the deviation. In several countries, central-bank officials are subject to public hearings in the Parliament where monetary policy is scrutinized; and in several countries, monetary policy is regularly or occasionally subject to extensive

reviews by independent experts (for instance, New Zealand, the UK, Norway, and Sweden).

So far, since its inception in the early 1990s, inflation targeting has been a considerable *success*, as measured by the stability of inflation and the stability of the real economy. There is no evidence that inflation targeting has been detrimental to growth, productivity, employment, or other measures of economic performance. The success is both absolute and relative to alternative monetary-policy strategies, such as exchange-rate targeting or money-growth targeting. No country has so far abandoned inflation targeting after adopting it, or even expressed any regrets. For both industrial and non-industrial countries, inflation targeting has proved to be a most flexible and resilient monetary-policy regime, and has succeeded in surviving a number of large shocks and disturbances. As of 2007, a long list of non-industrial countries were asking the International Monetary Fund for assistance in introducing inflation targeting. Although inflation targeting has been an unqualified success in all the small- and medium-sized industrial countries that have introduced it, the United States, the eurozone and Japan have not yet adopted all the explicit characteristics of inflation targeting, but they are all moving in that direction. Reservations about inflation targeting have mainly suggested that it might give too much weight on inflation stabilization to the detriment of the stability of the real economy or other possible monetary-policy objectives; the fact that real-world inflation targeting is flexible rather than strict and the empirical success of inflation targeting in the countries where it has been implemented seem to confound those reservations.

A possible alternative to inflation targeting is *money-growth targeting,* whereby the central bank has an explicit target for the growth of the money supply. Money-growth targeting has been tried in several countries but been abandoned, since practical experience has consistently shown that the relation between money growth and inflation is too unstable and unreliable for money-growth targeting to provide successful inflation stabilization. Although Germany's Bundesbank paid lip service to money-growth targeting for many years, it often deliberately missed its money-growth target in order to achieve its inflation target, and is therefore arguably better described as an implicit inflation targeter. Many small and medium-sized countries have tried exchange-rate targeting in the form of a *fixed exchange rate,* that is, fixing the exchange rate relative to a centre country with an independent monetary policy. For several reasons, including increased international capital flows and difficulties defending misaligned fixed exchange rates against speculative attacks, fixed exchange rates have become less viable and less successful in stabilizing inflation. This has led many countries to instead pursue inflation targeting with flexible exchange rates.

A current much-debated issue concerning the further development of inflation targeting is the appropriate *assumption about the instrument-rate path* that underlies the forecasts of inflation and other target variables and the *information provided about future policy actions.* Traditionally, inflation-targeting central banks have assumed a constant interest rate underlying its inflation forecasts, with the implication that a constant-interest-rate inflation forecasts that overshoots (undershoots) the inflation target at some horizon such as two years indicates that the instrument rate needs to increased (decreased). Increasingly, central banks have become aware of a number of serious problems with the assumption of constant interest rates. These problems include that the assumption may often be unrealistic and therefore imply biased forecasts, imply either explosive or indeterminate behaviour of standard models of the transmission mechanism of monetary policy, and on closer scrutiny be shown to combine inconsistent inputs in the forecasting process (such as some inputs such as asset prices that are conditional on market expectations of future interest rates rather than constant interest rates) and therefore produce inconsistent and difficult-to-interpret forecasts. Some central banks have moved to an instrument-rate assumption equal to market expectations at some recent date of future interest rates, as they can be extracted from the yield curve. This reduces the number of problems mentioned above but does not eliminate them. For

instance, the central bank may have a view about the appropriate future interest-rate path that differs from the market's view. A few central banks (notably in New Zealand, Norway, and Sweden – the last probably within the next few months) have moved to deciding on and announcing an optimal instrument-rate path; this approach solves all the above problems, is the most consistent way of implementing inflation targeting, and provides the best information for the private sector. The practice of deciding on and announcing optimal instrument-rate paths is now likely to be gradually adopted by other central banks in other countries, in spite of being considered more or less impossible, or even dangerous, only a few years ago.

Another issue is whether flexible inflation targeting should eventually be transformed into flexible *price-level targeting*. Inflation targeting as practised implies that past deviations of inflation from target are not undone. This introduces a unit root in the price level and makes the price level non-stationary. That is, the conditional variance of the future price level increases without bound with the horizon. In spite of this, inflation targeting with a low inflation rate is referred to as 'price stability'. An alternative monetary-policy regime would be 'price-level targeting', where the objective is to stabilize the price level around a price-level target. That price-level target need not be constant but could follow a deterministic path corresponding to a steady inflation of two per cent, for instance. Stability of the price level around such a price-level target would imply that the price level becomes trend stationary, that is, the conditional variance of the price level becomes constant and independent of the horizon. One benefit of this compared with inflation targeting is that long-run uncertainty about the price level is smaller. Another benefit is that, if the price level falls below a credible price-level target, inflation expectations would rise and reduce the real interest rate even if the nominal interest rate is unchanged. The reduced real interest rate would stimulate the economy and bring the price level back to the target. Thus, price-level targeting may imply some automatic stabilization. This may be highly desirable, especially in situations when the zero lower bound on nominal interest rates is binding, the nominal interest rate cannot be further reduced, and the economy is in a liquidity trap, as has been the case for several years until recently in Japan. Whether price-level targeting would have any negative effects on the real economy remains a topic for current debate and research.

See Also

- ▶ Central Bank Independence
- ▶ Inflation
- ▶ Inflation Dynamics
- ▶ Inflation Measurement

Bibliography

Roger, S., and M. Stone. 2005. On target? The international experience with achieving inflation targets. Working paper no. 05/163. Washington, DC: International Monetary Fund.

Svensson, L.E.O. 2002. Monetary policy and real stabilization. In *Rethinking stabilization policy: A symposium sponsored by the Federal Reserve Bank of Kansas City.* Kansas City: Federal Reserve Bank for Kansas City.

Svensson, L.E.O. 2007. Optimal inflation targeting: further developments of inflation targeting. In *Monetary policy under inflation targeting*, ed. F. Mishkin and K. Schmidt-Hebbel. Santiago: Banco Central de Chile.

Woodford, M. 2005. Central-bank communication and policy effectiveness. In *The Greenspan Era: Lessons for the future – A symposium sponsored by the Federal Reserve Bank of Kansas City.* Kansas City: Federal Reserve Bank for Kansas City.

Inflationary Gap

David Vines

This term originates from the analysis of inflation put forward by Keynes in *How to Pay for the War* (1940). If there is a gap between the level of aggregate demand for goods and services and the quantity of available supply, then this will cause inflation.

The 'inflation gap' has been used as a basis for straightforward demand pull theories of inflation. But in *How to Pay for the War* Keynes used his concept of the inflation gap to build a strikingly novel theory of the inflationary process in the UK during World War I, which foreshadowed both the demand pull and the cost push concepts developed later, and which also made a particular use of the effects of inflation on income distribution. First of all, Keynes embodied the assumption of flexible rather than administered prices of produced goods and services; any excess demand associated with the inflation gap causes the prices of goods to rise relative to their costs of production, to the extent necessary to choke off the excess demand. The second distinctive part of the theory is as follows. The rise in prices of goods will be effective in choking off the excess demand – i.e. it will close the inflation gap – for the following two reasons. First, if there is a tax on profits bigger than any tax on wages then profit incomes will leak out of circulation to the government thereby diminishing the consumption of rentiers. Second, if the propensity to consume out of profit incomes is lower than that out of wage incomes – as assumed by Keynes – then a redistribution away from wages will, of itself, lower the propensity to consume. Notice that, the fall in real wages required to close the inflation gap may be large, since it depends upon the existence of the profits tax or upon the difference between the two consumption propensities. The third distinctive element in Keynes's theory is that the resulting reduction of real wages will cause pressure for an increase in money-wages. But if the inflation gap is not to reappear, then the prices of goods must run ahead of the increases in money-wages again. The speed of the inflation depends upon the lag with which wages chase prices, and inflation accelerates if this lag shortens.

The novelty of Keynes theory may be apparent when it is realised that the dominant inflation theory of the time derived from the quantity theory of money, in which the rate of inflation is determined by the rate of monetary growth (Keynes 1940). Keynes's own theory has remained popular with Latin American structuralist writers; the inflation gap originates both because of chronic supply side ('structural') problems and because of a failure of policy to control demand, and changes in the distribution of income away from wages are thought to play an important role in stabilizing the overall process, as in Keynes's theory. (See Kaldor 1964; Cardoso 1980, presents a formal model but without the stabilizing role of lower real wages on demand.)

Orthodox demand pull theories of inflation see an inflation gap as leading to a rather different kind of inflationary process. Again they begin with an excess of aggregate demand over supply. In one version this excess demand will increase the prices of goods relative to wages, since firms are assumed to be price taking profit maximizers, who only increase supplies in the face of an extra demand if the profit margin per unit rises at the same time (Friedman 1968, 1975; Phelps 1970). In this case, wages again chase prices as wage earners begin to learn that real wages have fallen, and the process so far is very similar to that presented by Keynes. Or, in another version, the excess demand may not of itself raise prices until it percolates directly through to the labour market and causes money-wages to rise, subsequently inducing increases in prices, and then in turn inducing wages to chase the higher prices (Laidler and Parkin 1975). But in both of these more conventional demand pull theories, the extent of the inflation which ultimately emerges depends on how demand management policy (i.e. fiscal and monetary policy) responds to dampen down the inflation gap, by curtailing economic activity. Indeed, in the hands of monetarist economists, under the assumption of a constant velocity of circulation of money, the extent of the inflation comes to depend directly on the rate of growth of the money supply, the very view which Keynes had earlier criticized. Crucially, these orthodox theories do not, like Keynes, see changes in the distribution of income away from wages as serving the function of regulating the excess demand, and of thereby determining the severity of the inflationary process which emerges.

Keynes himself did not foreshadow another possibility, almost the reverse of what he analysed. This is that a cost push inflation process could be engendered by a social conflict over the distribution of income, quite independently of any inflationary stimulus coming from the inflation gap (see Rowthorn 1977; Meade 1982; and Marglin 1984).

See Also

- ▶ Demand Management
- ▶ Demand-Pull Inflation
- ▶ Fiscal Stance

Bibliography

Cardoso, E. 1980. Food supply and inflation. *Journal of Development Economics* 8: 269–284.
Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.
Friedman, M. 1975. Unemployment versus inflation? Occasional Paper No. 44. London: Institute of Economic Affairs. Reprinted in M. Friedman, *Lectures in price theory.* Chicago: Aldine, 1976.
Kaldor, N. 1964. Economic problems of Chile. Chapter 21 of N. Kaldor. In *Essays on economic policy*, vol. II. London: Duckworth.
Keynes, J.M. 1940. *How to pay for the war.* London: Macmillan. Reprinted in *Essays in persuasion; the collected writings of John Maynard Keynes*, vol. IX. London: Macmillan, 1972.
Laidler, D., and M. Parkin. 1975. Inflation: A survey. *Economic Journal* 85: 741–809.
Marglin, S.A. 1984. Growth distribution, and inflation: A centennial synthesis. *Cambridge Journal of Economics* 8: 115–144.
Meade, J.E. 1982. *Stagflation,* vol. 1: *Wage fixing*. London: George Allen & Unwin.
Phelps, E. (ed.). 1970. *Microeconomic foundations of employment and inflation theory.* London: Macmillan.
Rowthorn, R.E. 1977. Conflict, inflation and money. *Cambridge Journal of Economics* 1: 215–239.

Informal Economy

Keith Hart

Abstract

The informal economy or sector has become the preferred term for unregulated economic activities, in both rich and poor countries. Based on Weber's theory of rationalization, it was coined during the early 1970s in response to proliferating self- employment and casual labour in Third World cities. Now its range of reference is very wide, embracing everything from high-level political corruption to home improvement. The phenomenon is real enough and of some antiquity, but its definition remains elusive. Operating beyond the rules of bureaucracy, the informal economy may be understood dialectically as division, content, negation or residue.

Keywords

Bureaucracy; Capital accumulation; Economic development in the long run; Exploitation; Informal economy; Labour surplus economies; Poverty; Protection; Rational enterprise; Rule of law; Rural–urban migration; Self-employment; Weber, M.

JEL Classifications

C1

The term 'informal economy' became current in the 1970s as a label for economic activities that take place outside the framework of bureaucratic public and private sector establishments. It arose in response to the proliferation of self-employment and casual labour in Third World cities; but later the expression came to be used with reference to societies like Britain, where it competed with epithets of deindustrialization – the 'hidden', 'underground', 'black' economy, and so on.

The social phenomenon is real enough and of some antiquity. London's East End in the mid-19th century is a stark example of informal economic organization which rivals in scale any of today's tropical slum areas (Davis 2006). Nevertheless, the empirical referents of the 'informal economy' remain elusive, ranging as they do between the extremes of corrupt public finance in Congo and do-it-yourself in a London suburb. The intellectual history of the concept is clearer. It was provoked by the failure of prevalent economic models to address a large part of the world that they claimed to offer prescriptions for. Sociologists, anthropologists, geographers and historians have grasped the opportunity to embarrass economists by pointing out this deficiency. More remarkably, many economists, including employees of bureaucracies such as the World

Bank and the International Labour Organization (ILO), have identified the 'informal sector' as something they must deal with. Whereas once the effects of 'informality' were thought to be palliative, they are now often seen as a threat to legitimate businesses.

Some notable attempts have been made to document the economy of the streets. Henry Mayhew's investigations for the *Morning Chronicle* in the 1850s, published as *London Labour and the London Poor* (1861–2), are a classic source, as are Oscar Lewis's several accounts of the 'culture of poverty' (for example, *La Vida,* 1964). Very little of all this impinged on the world of development economists. The dualistic models of economic development that prevailed in the 1960s took their lead from W. Arthur Lewis's (1954) theory of development with unlimited supplies of labour, whereby underemployed rural workers migrated to find wage employment in a higher productivity urban economy.

In *Peddlers and Princes* (1963), Clifford Geertz identified two economic ideal types in a Javanese town. The majority were occupied in a street economy that he labelled 'bazaar-type'. Opposed to this was the 'firm-type' economy consisting largely of Western corporations that benefited from the protection of state law. These had *form* in Weber's (1981) sense of 'rational enterprise' based on calculation and the avoidance of risk. National bureaucracy lent these firms a measure of protection from competition, thereby allowing the systematic accumulation of capital. The 'bazaar', on the other hand, was individualistic and competitive, so that accumulation was well-nigh impossible. Geertz considered what it would take for a group of reform Muslim entrepreneurs to join the modern 'firm' economy. They were rational and calculating enough; but they were denied the institutional protection of state bureaucracy, which was the preserve of the existing corporations.

A decade later and in the context of growing unease over Third World urban unemployment, Keith Hart (1973, based on a conference paper of 1971) argued that the masses who were surplus to the requirements for wage labour in African cities were not 'unemployed' but rather were positively employed, even if often for erratic and low returns. He proposed that these activities be contrasted with the 'formal' economy of government and organized capitalism as 'informal income opportunities'. Moreover, he suggested that the aggregate inter-sectoral relationship between the two sources of employment might be of some significance for models of economic development in the long run. In particular, the informal economy might be a passive adjunct of growth originating elsewhere or its dynamism might be a crucial ingredient of economic transformation in some cases.

The dualism (formal–informal) and some of the thinking behind it received immediate publicity through its adoption in an influential ILO (1972) report on incomes and employment in Kenya, which elevated the 'informal sector' to the status of a major source for national development by the bootstraps, as it were. This was enough to encourage legions of researchers to adopt the term. Before long a substantial critique of the 'informal sector' concept had emerged. Marxists claimed that its proponents mystified the essentially regressive and exploitative nature of this economic zone, which they preferred to call 'petty commodity production'. The study of Third World urban poverty rapidly became a new segment of the academic division of labour; as a key term in its discourse, the informal economy attracted an unusual volume of debate (Bromley 1978). Later, sociologists applied the term to industrial societies (Pahl 1984).

Hernando De Soto argued that Peru was a mercantilist state whose overregulated and impenetrable national bureaucracy excluded the vast majority from effective participation in development. The latter were an entrepreneurial peasantry flocking in ever-larger numbers to the main cities. They were forced to operate informally, that is, outside the law, in sectors such as housing, trade and transport. Later, he portrayed poor countries like Peru as being trapped in a world economy dominated by the first industrial nations (De Soto 2000). Red tape is mainly an effect of a global regime that forces marginal states to adopt inappropriate institutional practices. The result is the same: migrants pile up in the cities and are

forced to work outside the law. Countries like the USA, which dominates this global financial bureaucracy, made the transition to modern capitalism by giving informal practices and decentralized violence full rein in their own development. Similar flexibility has to be shown today if the poor urban masses are to have a chance of joining global development on less unequal terms.

The idea of an 'informal economy' is entailed by the institutional effort to organize society along formal lines (Hart 2006). 'Form' is *the rule,* an idea of what ought to be universal in social life; and for most of the 20th century the dominant forms have been those of bureaucracy, particularly of national bureaucracy, since society has become identified to a large extent with nation states. This identity may now be weakening in the face of the neoliberal world economy and a digital revolution in communications. Popularity as a jargon word has not helped the informal economy acquire a measure of analytical precision. For many it is a convenient name for an unambiguous empirical phenomenon – what you find in the slums of Manila. Others refer to size (large-scale–small-scale), productivity (high–low), visibility (enumerated–unenumerated), pattern of rewards (wages–self-employment), market conditions (monopoly–competitive) and much else. Hart (1973), like Geertz, explicitly derived his analysis from Weber's theory of rationalization. Much that goes on in developing countries today is only marginally the product of state regulation: it is thus 'informal' relative to the forms of publicly organized economic life. This is a qualitative distinction.

'Form' is the rule, the invariant in the variable. Idealist philosophers from Plato onwards thought the general idea of something was more real than the thing itself. The 'formal sector' is likewise an idea, a collection of people, things and activities that share an idea; but we should not mistake the idea for the reality that it partially identifies. What makes something 'formal' is its conformity with such an idea or rule. Thus formal dress in some societies means that the men will come dressed like penguins, but the women are free to wear something extravagant that suits them personally – they come as variegated butterflies. Formality endows a class of people with universal qualities, with being the same and equal. What makes dress 'informal' is therefore the absence of such a shared code. But informality is relative to the eye of the beholder. Any observer of an informally dressed crowd will notice that the clothing styles are not random. We might ask what these informal forms are and how to account for them. The world's ruling elite can be identified as 'the men in suits', because they choose to wear a style invented in the 1920s as an informal alternative to formal evening dress.

What the public and private sectors share is conformity to the rule of law at the national and increasingly international levels. How then might non-conformist economic activities, 'the informal economy', relate to this formal order? They may be related in any of four ways: as *division,* as *content,* as *negation* and as *residue.* The first two imply a positive relationship of interdependence, the third is antagonistic and the last relatively autonomous. The moral economy of capitalist societies is based on an attempt to keep separate impersonal and personal spheres of social life. The establishment of a formal public sphere entailed another based on domestic privacy. Most people, traditionally men more than women, divide themselves every day between production and consumption, paid and unpaid work, submission to impersonal rules in the office and the free play of personality at home. Money is the means whereby the two sides are brought together, so that their interaction is an endless process of separation and integration that I call *division.*

For any rule to be translated into human action, something else must be brought into play, such as personal judgement. So informality is built into bureaucratic forms as unspecified *content.* Take a chain of commodities from their production by a transnational corporation to their final consumption in a Third World city. At several points invisible actors appear filling the gaps that the bureaucracy cannot handle directly, from the factories to the docks to the supermarkets and street traders who supply the cigarettes to smokers. Informal processes are indispensable to the trade, as variable content to the universal form. Of course, some of

these activities may break the law, through a breach of health and safety regulations, tax evasion, smuggling, the use of child labour, selling without a licence, and so on. The third way that informal activities relate to formal organization is thus as its *negation.* Rule breaking takes place both within bureaucracy and outside it; and so the informal is often illegal. The informalization of the world economy is to a large extent criminal and this includes white-collar crime.

The fourth category is not so obviously related to the formal order as the rest. Some 'informal' activities exist parallel to it, as *residue.* They are just separate from the bureaucracy. It would be stretching the logic of the formal–informal pair to include peasant economy, traditional institutions and much else besides within the rubric of the 'informal'. Yet the social forms endemic to these often shape informal economic practices.

It is inconsistent to claim that the urban poor have an informal economy but their rich masters do not; or that the Third World has an informal sector but not the industrialized West. As long as there is formal economic analysis and the *partial* institutionalization of economies around the globe along capitalist lines, there will be a need for some such remedial concept as the informal economy. Its application to concrete conditions is stimulated by palpable discrepancies between prevalent models and observed realities. Such a discrepancy provoked the emergence of the concept in the 1970s, when Third World economies bore the brunt of the depression that marked the end of the West's post-war miracle. Later the accelerating decline of the British economy encouraged some social scientists to adopt the term there. The common strand is the growing inability of modern states to control the wider economic environment that sustains them. Hence the need for a dualistic model, such as that offered by the 'informal economy' concept.

See Also

- ▶ Development Economics
- ▶ Economic Anthropology
- ▶ Labour Surplus Economies

Bibliography

Bromley, R., ed. 1978. The urban informal sector: Critical perspectives. *World Development* 6(9–10).

Davis, M. 2006. *Planet of slums*. New York: Verso.

De Soto, H. 2000. *The mystery of capital: Why capitalism triumphs in the west and fails everywhere else*. London: Bantam.

Geertz, C. 1963. *Peddlers and princes*. Chicago: University of Chicago Press.

Hart, K. 1973. Informal income opportunities and urban employment in Ghana. *Journal of Modern African Studies* 11: 61–89.

Hart, K. 2006. Bureaucratic form and the informal economy. In *Linking the formal and informal economy: Concepts and policies*, ed. B. Guha-Khasnobis, R. Kanbur, and E. Ostrom. Oxford: Oxford University Press.

ILO (International Labour Organization). 1972. *Incomes, employment and equality in Kenya*. Geneva: ILO.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economics and Social Studies* 22 (2): 139–191.

Lewis, O. 1964. *La Vida: A Puerto Rican family in the culture of poverty – San Juan and New York*. New York: Random House.

Mayhew, H. 1861–2. *London labour and the London poor,* 4 vols. London.

Pahl, R. 1984. *Divisions of labour*. Oxford: Basil Blackwell.

Weber, M. 1981. *General economic history*. New Brunswick: Transaction Publishers.

Information Aggregation and Prices

James Jordan

Abstract

Economists commonly interpret market-clearing prices as the signals that competitive markets transmit to economic agents to facilitate the efficient allocation of resources. Informational decentralization theory formalizes this interpretation by characterizing the market mechanism as the unique decentralized mechanism that achieves efficient allocation with the minimal required communication. Rational expectations equilibrium theory formalizes a different aspect of the interpretation, showing

that markets transmit to each trader all of the decision-relevant information in the market.

Keywords

Cobb–Douglas functions; Efficient markets hypothesis; First Fundamental Welfare Theorem; Full communication equilibrium; General equilibrium under uncertainty; Hayek, F.; Information aggregation and prices; Informational decentralization; Rational expectations equilibrium

JEL Classifications

D8; G1

Market-clearing prices aggregate decision-relevant information that is initially dispersed throughout the economic environment. Hayek (1945) asserts that competitive markets economize on the communication needed to achieve efficient allocations by embedding in prices all that any individual needs to know about the rest of the economy. During the 1970s and 80s, Hayek's famous assertion was interpreted and formalized by two distinct literatures in economic theory. Hurwicz's (1960) model of decentralized allocation mechanisms stimulated a literature on *informational decentralization theory* that led to the characterization of the market as the unique informationally decentralized allocation mechanism that minimizes the communication needed to achieve Pareto-efficient allocations. This result verifies Hayek's assertion, although the minimal message communicated by the market mechanism necessarily includes the market clearing trades as well as prices.

Hayek's assertion has a second connotation in financial asset markets and other markets involving trade under uncertainty, where the dispersed information may be of direct interest to all traders. In this setting Hayek's assertion can be interpreted as a version of the *strong form* of the *efficient market hypothesis* (for example, Fama 1970), which states that market prices constitute a sufficient statistic for all of the decision-relevant information in the market. Adding *rational expectations* (Muth 1961) to models of general equilibrium under uncertainty makes it possible to formalize Hayek's assertion as follows. If all information in the market is directly communicated to all traders, the resulting equilibrium is also a rational expectations equilibrium. That is, each trader would find the market-clearing prices statistically sufficient for all of the decision-relevant private information of others. This version of Hayek's assertion is also verified, again provided that the market clearing trades are added to the prices in forming the sufficient statistic. There are also interesting cases in which the prices alone form a sufficient statistic.

Informational Decentralization

Mount and Reiter (1974) formalized the general model of informationally decentralized allocation as follows. There are n agents, indexed $1 \leq i \leq n$. The private information, or environment, of agent i is an element e^i of a set E^i. The set of economic environments is the Cartesian product $E = E^1 \times \ldots \times E^n$, with generic element $e = (e^1,\ldots,e^n)$. Let Y denote a set of outcomes or allocations. The desired performance is modelled by a *performance correspondence* $g : E \rightarrow\!\!\!\!\rightarrow Y$, which associates with each environment e a set of desired allocations $g(e)$, where the double arrow is used to denote that g is set-valued. The communication among agents is embodied in messages $m \in M$, where M is the *message space*. Each message is associated with an allocation via an *outcome* function $h : M \rightarrow Y$. Finally, each environment is associated with a set of messages via a *message correspondence* $\mu : E \rightarrow\!\!\!\!\rightarrow M$. An allocation mechanism (μ,M,h) *realizes* the performance correspondence g if for all $e \in E$,

$$\mu(e) \neq \varnothing; \tag{1}$$

and

$$h(m) \in g(e) \text{ for all } m \in \mu(e). \tag{2}$$

An allocation mechanism (μ,M,h) is *informationally decentralized* if there are individual message correspondences $\mu^i : E^i \rightarrow\!\!\!\!\rightarrow M$ such that for each $e \in E$,

$$\mu(e) = \cap_{i=1}^{n} \mu^i(e^i). \quad (3)$$

This rather abstract definition can be interpreted by imagining that a computer chooses a message m at random and displays it to all agents. If for some agent i, $m \notin u^i(e^i)$ then agent i vetoes this message and another is chosen at random, until a message m is found that is accepted by all agents, that is, $m \in \cap_i \mu^i(e^i)$, at which point the allocation $h(m)$ is chosen. The message m embodies the communication needed to achieve the allocation $h(m) \in g(e)$, because each agent i makes the veto/ accept decision based only on e^i and m. An informationally decentralized allocation mechanism (μ,M,h) that realizes a given g is *informationally efficient* if there is no other such mechanism that uses a message space that is smaller, in an appropriate sense, than M. If E and Y are finite, then the cardinality of M is the appropriate measure of size. In this case, the minimum required size agrees with the *communication complexity* of g, as that measure is defined in the computer science literature (for example, Karchmer 1989). In models with continua of environments and allocations, it is more common to require that M be a manifold and interpret its dimension as its size. In this case, informational efficiency has the interpretation of using messages with the minimum number of real variables.

The application of this model to competitive markets is direct. Let E denote the set of pure exchange environments with l commodities. Assume that each trader's consumption set is the non-negative orthant R^l_+, so that trader i's private information is $e^i = (u^i,\omega^i)$, where $u^i : \mathrm{R}^l_+ \to \mathrm{R}$ is trader i's utility function and ω^i is an initial endowment bundle in R^l_+. Let Y be the set of all net trades that balance in the aggregate, $Y = \{y = (y^1, \ldots, y^n) \in \mathrm{R}^{nl} : \sum_i y^i = 0\}$. The desired allocations are simply the Pareto efficient allocations that are also non-coercive, in the sense that traders are not forced below the utility level of their initial endowments. Formally, define $g : E \to\!\!\!\!\to Y$ as $g(e) = \{y \in Y : (\omega^1 + y^1, \ldots, \omega^n + y^n)$ is a Pareto-efficient allocation satisfying $u^i(\omega^i + y^i) \geq u^i(\omega^i)$ for all $i\}$. Non-coerciveness, which is sometimes called 'individual rationality', excludes the possibility of achieving Pareto efficiency by giving everything to one trader.

The *competitive allocation mechanism* is defined as follows. Let P denote the interior of the unit simplex in R^l_+, and define the *competitive message space* $M_c = \{(p, y) \in P \times Y : py^j = 0$ for all $i\}$ The outcome function $hc : Mc \to Y$ is the projection $h_c(p, y) = y$, and the individual message correspondence $\mu^i_c : E^i \to\!\!\!\!\to M_c$ is defined as $\mu^i_c(u^i, \omega^i) = \{(p, y) \in M_c. : \omega^i + y^i \in \mathrm{R}^l_+$ and $u^i(\omega^i + y^i) \geq u^i(x^i)$ for all $x_i \in \mathrm{R}^l_+$ satisfying $px^i \leq p\omega^i\}$. In effect, $\mu^i_c(e^i)$ is trader i's offer curve, and the *competitive message correspondence* μ_c, defined as $\mu_c(e) = \cap_i \mu^i_c(e^i)$, is the intersection of the offer curves.

The competitive allocation mechanism (μ_c,M_c, h_c) is informationally decentralized by construction. If the sets E^i are restricted by conventional assumptions (for example, utility functions are continuous, quasi-concave and strictly increasing, and endowments are strictly positive) then $\mu_c(e)$ is equal to the nonempty set of competitive equilibria for e, and the First Fundamental Welfare Theorem implies that the competitive allocation mechanism realizes g.

The informational efficiency of the competitive allocation mechanism was established by Hurwicz (1977) and Mount and Reiter (1974). The dimension of the competitive message space, M_c, is $n(l\text{-}1)$, so the informational efficiency of (μ_c,M_c,h_c) means that any other allocation mechanism (μ,M,h) which is informationally decentralized and realizes g must have $\dim M \geq n(l - 1)$. This requires imposing sufficient mathematical regularity on (μ,M,h) so that dim M is well-defined and μ cannot behave as a Peano curve, encoding multi-dimensional information into the unit interval, for example. In particular, it is sufficient to require that M be (homeomorphic to) an open subset of a Euclidean space and that, on the set of exchange environments in which all traders have Cobb–Douglas utility *functions*, the correspondence μ admits a continuous selection on some open subset (on Cobb–Douglas environments, μ_c is itself single-valued and continuous). More general conditions are given by Mount and

Reiter (1974), where the non-coerciveness requirement on g is also relaxed to require merely that for Cobb–Douglas environments, $g(e)$ includes only interior Pareto-efficient allocations. This excludes the mechanism that gives everything to trader 1 by using a message space of dimension $(n - 1)$l to enable traders 2,..., n to communicate their endowments.

The informational efficiency of competitive markets leaves open the possibility that other allocation mechanisms are also informationally efficient. This possibility is excluded by Jordan (1982b), albeit under stronger mathematical regularity conditions. In particular, the message correspondence μ is required to be single-valued and continuous on Cobb–Douglas environments, as opposed to merely having a local continuous selection. The non-coerciveness assumption is also much less dispensable for this result.

The informational decentralization literature verifies Hayek's assertion by characterizing the market mechanism as the unique informationally efficient mechanism that achieves non-coercive Pareto-efficient allocations. However, the competitive message is more than just the 1 - 1 relative prices that are the focus of Hayek's insight. The realization of non-coercive Pareto-efficient allocations requires $n(1–1)$-dimensional messages because of the need to communicate the equilibrium trades as well as the prices.

Rational Expectations Equilibrium

A simple version of the rational expectations equilibrium model can be described as follows. Before trade, each trader i observes her endowment, $\omega^i \in \mathrm{R}^1_+$, and a private signal, z_i, which is jointly distributed with the future state s, which is common to all traders, that determines her utility function $u^i(\cdot, s) : \mathrm{R}^1_+ \to \mathrm{R}$. Assume for simplicity that there is only a finite number of possible private signal values, each of which has positive probability. In a rational expectations equilibrium, each trader i maximizes her expected utility conditional on her private signal z^i and any endogenous market variables she observes. To formulate the information aggregation condition, suppose that all private signals were publicly observable. Then every trader would observe all of the information in the market, so there would be no need to infer information from market variables. The appropriate equilibrium concept would be a *full communication equilibrium* (FCE), defined as associating with each profile of signals $z = (z^i)_i$, a price vector $p(z)$ and net trades $y(z) = (y^i(z))_i$ satisfying, for each z,

$$\sum\nolimits_i y^i(z) = 0, \tag{4}$$

and for each trader i,

$$\begin{aligned} &\omega^i + y^i(z) \text{ maximizes } E\{u^i(x^i,\cdot)\,|\,z\} \\ &\quad \text{subject to } p(z)\ x^i \le p(z)\omega^i, \end{aligned} \tag{5}$$

where the expectation is taken over s with respect to the conditional distribution over s given z. Thus, an FCE allocation is an allocation that would result if every trader possessed all of the information in the market. The information aggregation question is thus whether an FCE can be supported if traders are given only their private information and, for example, the equilibrium price vector. More precisely, does an FCE $(p(\cdot), (y^i(\cdot))_i)$ also satisfy

$$\begin{aligned} &\omega^i + y^i(z) \text{ maximizes } E\{u^i(x^i,\cdot)\,|\,z^i, p(z)\} \\ &\quad \text{subject to } p(z)x^i \le p(z)\omega^i, \end{aligned} \tag{6}$$

for each z and each trader i? Functions $p(\cdot)$, $(y^i(\cdot))_i$ that satisfy Eqs. (4) and (6) constitute a *rational expectations equilibrium* (REE). Thus the information aggregation question is whether an FCE is also an REE.

Kreps (1977) provides a simple example showing not only that the answer is 'no', but that an REE can easily fail to exist. In the Kreps example, there are two traders, two commodities and two equiprobable states of the world, $s \in \{a, b\}$ The traders' endowments are $\omega^1 = \omega^2 = (3, 3)$ and their state-dependent utility functions are given by

$$\begin{aligned} u^1(x,a) &= 2lnx_1 + lnx_2,\ u^1(x,b) \\ &= lnx_1 + 2lnx_2, \end{aligned}$$

$u^2(x,a) = lnx_1 + 2lnx_2$, $u^2(x,b) = 2lnx_1 + lnx_2$.

Trader 1's signal is the state itself, $z^1 = s$, so trader 1 is fully informed; and trader 2's signal is a constant, $z^2 = \overline{z}^2$, so trader 2 is uninformed. Suppose by way of contradiction that $p(\cdot)$ is an REE price function. There are two possible cases. If $p(a,\overline{z}^2) \neq p(b,\overline{z}^2)$, then the price reveals the state to trader 2, and both traders are fully informed. However, it is easily seen that the (fully informed) environments $(\omega^i, u^i(\cdot, a))_i$ and $(\omega^i, u^i(\cdot, b))_i$ have the same unique equilibrium price $p = \left(\frac{1}{2}, \frac{1}{2}\right)$ There remains only the case that $p(a,\overline{z}^2) = p(b,\overline{z}^2)$, in which trader 1 is fully informed but trader 2 remains uninformed. However, the exchange environments $(\omega^1, u^1(\cdot,a)), (\omega^2, \frac{1}{2}u^2(\cdot,a) + \frac{1}{2}u^2(\cdot,b)))$ and $(\omega^1, u^1(\cdot,b)), (\omega^2, \frac{1}{2}u^2(\cdot,a) + \frac{1}{2}u^2(\cdot,b)))$ have unique and distinct equilibrium prices, thus eliminating this case as well. Thus market prices cannot always aggregate all the information in the market, and the use of market prices as information signals as well as rates of exchange can prevent even the existence of equilibrium.

If traders condition their expectations on the entire competitive message, however, full information aggregation occurs. In fact, for each trader i, conditioning on $p(z)$ and $y^i(z)$ is enough. More precisely, every FCE also satisfies Eq. (6)

$$\omega^i + y^i(z) \text{ maximizes } E\left\{u^i\left(x^i,\cdot\right) \middle| p(z), y^i(z)\right\}$$
$$\text{subject to } p(z)x^i \leq p(z)\omega^i, \qquad (7)$$

for each z and every trader i. It follows from the FCE property Eq. (5) that for any observed p and y^i, $\omega^i + y^i$ maximizes $E\{u^i(x^i, \cdot)| z\}$ subject to $px^i \leq p\omega^i$ for every z in the observed event $\{z : (p(z), y^i(z)) = (p, y^i)\}$. Thus the conditional expected utility function in Eq. (7) is a convex combination of expected utility functions, each of which is maximized by $\omega^i + y^i(z)$ subject to $p(z)x^i \leq p(z)\omega^i$. Therefore, the convex combination has the same maximum at the same constraint. Moreover, $(p(z), y^i(z))$ is the minimal market data needed for full information aggregation. Jordan (1982a) shows that, if traders condition expectations on non-constant functions of the competitive message $(p(z), (y^i(z))_i)$, then each trader's data must be sufficient to reveal $(p(z), y^i(z))$ to each trader i, not only to ensure full information aggregation but even to avoid examples of nonexistence of equilibrium.

The simultaneous determination of expectations, prices and trades in a rational expectations equilibrium begs the question of how the private information z becomes embedded in the prices and trades, from which each trader can infer the private information of others. A dynamic interpretation is given by Jordan (1982c). Suppose that, initially, each trader i conditions expectations on the private signal z^i alone. This leads to an initial equilibrium price vector $p_1(z)$, but suppose that, before the equilibrium trades are executed, traders update their expectations using the information revealed by $p_1(z)$. This leads to a second equilibrium price vector $p_2(z)$, and so on, until a price vector $p_T(z)$ is reached that reveals no new information that changes any trader's demand. This process may fail to reveal all decision-relevant information to every trader. For example, if the only trader with a non-constant signal has state-independent preferences, no information will be revealed and the process will terminate at the first step. However, for each trader i, the final price and net trade $\left(p_T(z), y^i_T(z)\right)$ is a sufficient statistic for all of the decision-relevant information trader i has learned from z^i and the temporary equilibrium prices. In this sense, the final prices and net trades summarize all of the private information revealed by prices along the temporary equilibrium path. The sequence of temporary equilibria is virtual in the sense that the temporary equilibrium trades are never executed. If they were, expectations of interim capital gains and losses could lead to nonexistence of temporary equilibrium, which is shown by an example in Jordan (1982c).

The Kreps (1977) example described above shows that prices alone cannot always support full information aggregation. However, the example is non-generic in the sense that a slight

perturbation of the state-dependent utility functions can make the full communication equilibrium prices different in the two states, resulting in full revelation. Radner (1979) develops a financial asset market model in which the FCE price function p(·) is generically 1–1. In Radner's model, the set of future states and current signal values are both finite. Each future state corresponds to a vector of values for the assets that are currently traded. Each signal z is associated with a conditional probability vector over the future states, and thus the future asset values. Radner shows that the set of signal-conditional probability arrays that give rise to FCE price functions that are not 1 - 1 is a closed nowhere dense set of Lebesgue measure zero. This line of research was greatly extended in a series of papers by Allen (see Allen and Jordan 1998, for a more detailed survey). Let Z denote the range of possible signal values z, and suppose that the relation between the signal z and conditional expected utility functions is sufficiently regular that the set of full communication environments $(E\{u^i(\cdot)|\ z\},\ \omega^i)_i$ has dimension no larger than the dimension of Z. If Z is finite, as in Radner's model, both sets have dimension zero.

Allen (1981) shows that, if $dimZ < \frac{1}{2}dimP$, then an FCE price function p(·) is generically 1-1. Allen (1982b) shows that, if $dimZ < dimP$, then an FCE price function is generically 1 - 1 except on a subset of Z having Lebesgue measure zero. This implies that, if the probability distribution over signals has a density function on Z, then an FCE price function is 1 - 1 on a set of signals having probability one, so that prices are again fully informative. The dimensional inequality is crucial. Jordan and Radner (1982) provide a robust example of the nonexistence of price-conditional rational expectations equilibrium with $dimZ = dimP = 1$.

Most of the results described above do not substantially restrict the way in which traders' preferences depend on the unknown future state of the world. Financial asset market models, in contrast, typically involve special kinds of state-dependent preferences that give rise to some interesting cases in which market prices are fully revealing. The earliest full revelation result was obtained by Jerry Green (1973) in an Arrow-securities markets model. In this case, the securities traded are wealth claims contingent on each future state, and traders have private signals about the probability distribution over the future states. Green (1973) shows that the derivative of market excess demand with respect to the state probabilities has a dominant diagonal property that ensures that the function from the full communication probabilities to the FCE price vector is 1-1. Grossman (1981) generalizes Green's model to obtain the full revelation of decision-relevant information even when the FCE prices are not 1-1. However, Green (1977) shows that, if 'noise' is included in the environment in the form of random endowments, rational expectations equilibrium can fail to exist.

The Green–Grossman full revelation result depends on the completeness of the securities markets. In the absence of complete markets, full revelation through prices can be obtained under restrictions on the nature of the uncertainty or on traders' utility-of-wealth functions. Grossman (1978) considers a model with a single riskless asset and several risky assets. The future values of the risky assets have a joint normal distribution. Traders have private signals about the mean of this distribution, but the covariance matrix is fixed. Grossman (1978) shows that, if traders' utility-of-wealth functions exhibit non-increasing absolute risk-aversion, then the FCE price vector is a 1 - 1 function of the full communication mean, and thus reveals all decision-relevant information. The same asset markets are studied by Jordan (1983), but arbitrary small perturbations are allowed in traders' endowments and the joint probability distribution over private signals and future risky asset values. In this case, if the number of private signal variables exceeds the number of risky assets ($dimZ > dimP$), full revelation by prices is assured only for three special classes of utility-of-wealth functions: linear, exponential, and constant relative risk aversion with the same constant for all traders.

I

The full revelation of private information by the market seems inconsistent with the acquisition of costly private information. For this reason, Grossman and Stiglitz (1980) introduced a financial asset market model, generalized by Hellwig (1980), which has a price-conditional rational expectations equilibrium that is only partially revealing. This model assumes that traders have exponential utility, and that future risky asset values are normally distributed. Full revelation is prevented by adding noise to the model in the form of randomness in the aggregate supply of the risky asset. Unfortunately, the existence of rational expectations depends on the special parametric assumptions of the model. Allen (1982a, 1985a, b) and Anderson and Sonnenschein (1982) develop general models of partially revealing approximate rational expectations equilibria, but the rational expectations equilibrium literature has not produced a general model of partially revealing equilibrium.

See Also

► Efficient Markets Hypothesis

Bibliography

Allen, B. 1981. Generic existence of completely revealing equilibria for economies with uncertainty when prices convey information. *Econometrica* 49: 1173–1199.

Allen, B. 1982a. Approximate equilibria in microeconomic rational expectations models. *Journal of Economic Theory* 26: 244–260.

Allen, B. 1982b. Strict rational expectations equilibria with diffuseness. *Journal of Economic Theory* 27: 20–46.

Allen, B. 1985a. The existence of rational expectations equilibria in a large economy with noisy price observations. *Journal of Mathematical Economics* 14: 67–103.

Allen, B. 1985b. The existence of fully rational expectations approximate equilibria with noisy price observations. *Journal of Economic Theory* 37: 213–253.

Allen, B., and J. Jordan. 1998. The existence of rational expectations equilibrium: A retrospective. In *Organizations with Incomplete Information*, ed. M. Majumdar. Cambridge: Cambridge University Press.

Anderson, R., and H. Sonnenschein. 1982. On the existence of rational expectations equilibrium. *Journal of Economic Theory* 26: 261–278.

Fama, E. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417.

Green, J. 1973. Information, efficiency and equilibrium. Discussion Paper No. 284. Cambridge, MA: Harvard Institute of Economic Research, Harvard University.

Green, J. 1977. The non-existence of informational equilibria. *Review of Economic Studies* 44: 451–463.

Grossman, S. 1978. Further results on the informational efficiency of competitive stock markets. *Journal of Economic Theory* 18: 81–101.

Grossman, S. 1981. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 48: 541–559.

Grossman, S., and J. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 35: 393–408.

Hayek, F. 1945. The use of knowledge in society. *American Economic Review* 35: 519–530.

Hellwig, M. 1980. On the aggregation of information in competitive markets. *Journal of Economic Theory* 22: 477–498.

Hurwicz, L. 1960. Optimality and informational efficiency in resource allocation processes. In *Mathematical Methods in the Social Sciences*, ed. K. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

Hurwicz, L. 1977. On the dimensional requirements of informationally decentralized Pareto-satisfactory processes. In *Studies in Resource Allocation Process*, ed. K. Arrow and L. Hurwicz. Cambridge: Cambridge University Press.

Jordan, J. 1982a. Admissible market data structures: A complete characterization. *Journal of Economic Theory* 28: 19–31.

Jordan, J. 1982b. The competitive allocation process is informationally efficient uniquely. *Journal of Economic Theory* 28: 1–18.

Jordan, J. 1982c. A dynamic model of expectations equilibrium. *Journal of Economic Theory* 28: 235–254.

Jordan, J. 1983. On the efficient markets hypothesis. *Econometrica* 51: 1325–1344.

Jordan, J., and R. Radner. 1982. Rational expectations in microeconomic models: An overview. *Journal of Economic Theory* 26: 201–223.

Karchmer, M. 1989. *Communication Complexity: A New Approach to Circuit Depth*. Cambridge, MA: MIT Press.

Kreps, D. 1977. A note on 'fulfilled expectations' equilibria. *Journal of Economic Theory* 14: 32–43.

Mount, K., and S. Reiter. 1974. The informational size of message spaces. *Journal of Economic Theory* 8: 161–192.

Muth, J. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.

Radner, R. 1979. Rational expectations equilibrium: Generic existence and the information revealed by prices. *Econometrica* 47: 665–678.

Information Cascade Experiments

Lisa Anderson and Charles A. Holt

Keywords

Bayes' rule; Imitation; Imperfect information; Information cascade experiments; Private information

JEL Classifications

C9

Cascade experiments test the theory that conformity can result from individuals receiving private imperfect information and making public decisions in a sequence (see ▶ Information Cascades).

Cascade theories provide a rational explanation for imitation even when people receive different private information. If a person gathers additional information by observing others' decisions, then a sequence of decisions that matches one alternative might be strong enough to outweigh that person's contrary private information. When the initial decisions in a sequence are correct, cascades can lead to better overall decision-making than private information alone. However, information cascades are problematic when the initial decision-makers in a queue receive incorrect information and convey it to others through their public (incorrect) decisions.

Anderson and Holt (1997) designed the first laboratory cascade experiment to test the theory described in Bikhchandani et al. (1992). Participants were shown two cups labelled A and B. Cup A contained two light marbles and one dark marble. Cup B contained two dark marbles and one light marble. A six-sided die was used to determine whether Cup A or Cup B was selected at the start of each decision-making round. The cups were equally likely to be selected by the die throw. Once a cup was selected, each person saw one private draw from the cup, with the marble being returned to the cup after each draw. Each participant made a public prediction about which cup (A or B) was being used for the draws in a randomly determined sequence that changed from round to round. Sessions included six decision-makers who were paid two dollars for a correct prediction and nothing otherwise for each of 15 rounds.

In any given round, if the first two public predictions matched (AA or BB) it was rational (based on Bayes' rule) for all subsequent decision-makers to follow, regardless of which marble they saw drawn from the cup (see ▶ Bayesian Statistics). Starting with prior probabilities of 1/2 for each cup, if the first decision-maker predicted cup A, others could rationally infer that he saw a light marble, since there were more light marbles than dark marbles in Cup A. With this new information, the probability of Cup A should have been updated to 2/3. If the second decision-maker predicted Cup A, others could infer that he also saw a light marble, and the probability of Cup A being used for the draws should have been updated to 4/5. Even if the third person observed a dark marble, it was still more likely that Cup A was being used for the draws, and a cascade should start with the third decision-maker. Alternatively, if the first two decision-makers cancelled each other out (AB or BA) and the next two matched, then a cascade could start with the fifth person in the sequence.

Cascades were possible, based on the private draws and the decision-making sequence, in about half the Anderson and Holt (1997) experiments and actually formed in about 70 per cent of these cases. Almost all the people who did not join rational cascades were following private information that conflicted with the cascade. This type of deviation is explained by cascade models with small amounts of noisy behaviour, as described in Anderson and Holt (1997) and Goeree et al. (2007), who showed that incorrect cascades are not likely to persist in experiments with long sequences of decisions.

From a policy perspective, cascades are a concern because they hide information, since the private information of cascade followers is not revealed by their decisions. Kübluer and Weizsäcker (2004) studied whether or not

people recognized the lack of information in conforming decisions by making participants pay a fee to see a private signal. In one version of their experiment, it was rational for only the first person in the sequence to purchase information, but the authors found that many people made irrational purchases. Some of this behaviour can be explained by a model with error, since it is rational to buy information if one cannot completely trust the quality of public decisions.

In addition to the studies discussed above, laboratory experiments have been used to test other variations of the seminal cascade theory including applications to voting (Hung and Plott 2001), investment (Alsopp and Hey 2000), markets (Drehmann et al. 2005; Cipriani and Guarino 2005) and advice-giving (Çelen et al. 2005).

See Also

- ▶ Bayesian Statistics
- ▶ Information Cascades

Bibliography

Alsopp, L., and J.D. Hey. 2000. Two experiments to test a model of herd behavior. *Experimental Economics* 3: 121–136.

Anderson, L.R., and C.A. Holt. 1997. Information cascades in the laboratory. *American Economic Review* 87: 847–862.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 1992. A theory of fads, fashion, custom and cultural change as information cascades. *Journal of Political Economy* 100: 992–1026.

Çelen, B., S. Kariv, and A. Schotter. 2005. Words speak louder than actions and improve welfare: An experimental test of advice and social learning. Working paper, Center for Experimental Social Science, New York University.

Cipriani, M., and A. Guarino. 2005. Herd behavior in a laboratory financial market. *American Economic Review* 95: 1227–1443.

Drehmann, M., J. Oechssler, and A. Roider. 2005. Herding and contrarian behavior in financial markets: An internet experiment. *American Economic Review* 95: 1203–1426.

Goeree, J.K., T.R. Palfrey, B.W. Rogers, and R.D. McKelvey. 2007. Self-correcting information cascades. *Review of Economic Studies* (forthcoming).

Hung, A.A., and C.R. Plott. 2001. Information cascades: Replication and an extension to majority rule and conformity-rewarding institutions. *American Economic Review* 91: 1508–1520.

Kübler, D., and G. Weizsäcker. 2004. Limited depth of reasoning and failure of cascade formation in the laboratory. *Review of Economic Studies* 71: 425–441.

Information Cascades

Sushil Bikhchandani, David Hirshleifer and Ivo Welch

Abstract

An information cascade occurs when individuals, having observed the actions and possibly payoffs of those ahead of them, take the same action regardless of their own information signals. Informational cascades may realize only a fraction of the potential gains from aggregating the diverse information of many individuals, which helps explain some otherwise puzzling aspects of human and animal behaviour. For example, why do individuals tend to converge on similar behaviour? Why is mass behaviour prone to error and fads? The theory of observational learning, and particularly of information cascades, has much to offer economics and other social sciences.

Keywords

Convergent behaviour; Herding; Imitation; Information aggregation; Information cascades and observational learning; Observational learning; Product life cycle; Signalling; Uncertainty

JEL Classification

D8; G1

An *information cascade* is a situation in which an individual makes a decision based on observation of others without regard to his own private information.

Social observers have long recognized that human beings have a deep-rooted proclivity to imitate. According to Machiavelli (1514, p. 152), 'Men nearly always follow the tracks made by others and proceed in their affairs by imitation.' Even animals imitate in choices of mate and territories. A common view among social scientists equates the conformity of individuals in large groups with irrationality – 'fads', 'mass psychology', or the 'madness of crowds'.

However, there has also been recent recognition of the benefits of social influence. For example, zoologists have argued that, despite its possible disadvantages, imitation is an evolutionary adaptation that has promoted survival over thousands of generations by allowing individuals to take advantage of the hard-won knowledge of others (Gibson and Hoglund 1992).

Nevertheless, as this article discusses, even when individuals are entirely rational, observational influence helps surprisingly little, leading to social outcomes that are inefficient and superficially may seem irrational. Irrationality undoubtedly affects social behaviour. Recent developments in the theory of observational learning, however, give reason to be sceptical about casual attributions of perverse social outcomes to irrational passions.

Why do people tend to 'herd' on similar actions? Why is mass behaviour prone to error and fads? The theory of observational learning helps explain some otherwise puzzling phenomena about human behaviour, and offers a vantage point for treating issues in economics and business strategy.

We call influence resulting from rational processing of information gained by observing others *observational learning* or *social learning*. Observational learning is only one of several possible causes of convergent behaviour. The simplest reason is that individuals can have identical beliefs and decision problems. Alternative reasons for conformity include positive payoff externalities, which lead to conventions such as driving on the right-hand side of the road; preference interactions, as with everyone desiring to wear the more 'fashionable' clothing as determined by what others are wearing; and sanctions against deviants, as with a dictator punishing opposition.

Among these theories, however, only observational learning explains why mass behaviour is error-prone, idiosyncratic, and often fragile in the sense that small shocks might lead to large shifts in behaviour. To understand how these effects arise, consider a sequence of rational individuals who take identical decisions under uncertainty. Each individual makes use of all relevant information – his own private signal and any inferences drawn from observing the choices of preceding individuals. As soon as the information gleaned from publicly observable choices of others is even slightly more informative than the individual's private signal, he imitates his immediate predecessor without regard to his private information. Therefore, this individual's choice is uninformative about his signal, and at that point an information cascade starts. His immediate successor finds herself in an identical position; she imitates him (her immediate predecessor) and ignores her private signal. Based on the information conveyed by the actions of the first few individuals – the ones not in a cascade – every succeeding individual takes the same action. This action may be an incorrect one, so even small shocks such as the possible arrival of a different type of individual or a little new information can overturn it. Thus, observational learning explains not only conformity but also rapid and short-lived fluctuations such as fads, fashions, booms and crashes.

The social outcome is highly error-prone because there is an information externality. If an individual selects an action that depends on his information signal, his action provides useful information to later decision-makers. However, it is in the self-interest of an individual in a cascade to ignore his signal; therefore, later individuals do not get the benefit of learning his private signal. Thus, the failure of individuals to take into account the welfare of later decision-makers leads to inefficient information aggregation.

This entry focuses on the situation where individuals with diverse private information learn by observing the actions of others or the consequences of these actions. (Previous surveys of this literature include Bikhchandani et al. 1998, and Chamley 2004.)

Observable Actions Versus Observable Signals

Consider a setting in which individuals choose an action in a chronological order. Each individual starts with some private information, obtains some information from predecessors, and then decides on a particular action. We consider two scenarios. In the *observable actions* scenario, individuals can observe the actions but not the signals (that is, private information) of their predecessors. As demonstrated below, cascades will arise in this model. We compare this with a benchmark *observable signals* scenario in which individuals can observe both the actions and the signals of predecessors. (See Welch 1992; Bikhchandani et al. 1992; Banerjee 1992.)

The main ideas are seen in the following simple example. Several risk-neutral individuals decide in sequence whether to *adopt* or *reject* a possible action. The payoff to adopting, V, is either 1 or -1 with equal probability; the payoff to rejecting is 0. In the absence of further information, the two alternatives are equally desirable. The order in which individuals decide is given and known to all.

Each individual's signal is either High (H) or Low (L). It is H with probability $p > 1 = 2$ if $V = 1$, and with probability $1-p$ if $V = -1$. Bayes' rule implies that, after observing one H, an individual's posterior probability that $V = 1$ is p; if instead one L is observed the probability that $V = 1$ is $1-p$. All private signals are identically distributed and independent conditional on V. Naturally, an individual's posterior belief about V also depends on information derived from predecessors. All this is common knowledge among the individuals.

In the observable signals scenario, each individual observes predecessors' information signals. As the pool of public information keeps increasing, later individuals will settle on the correct choice (adopt if $V = 1$, reject if $V = -1$) and thus behave alike.

Because actions reflect information, it is tempting to infer that, if only the actions of predecessors are observable, the public information set will also gradually improve until the true value is revealed almost perfectly. But that is not the case. In the observable actions case, individuals often converge fixedly on the same wrong action – that is, the choice that yields a lower payoff, *ex post*. Furthermore, behaviour is *idiosyncratic* in that the choices of a few early individuals determine the choices of all successors.

To return to our example, the first individual, Asterix, adopts if his signal is H and rejects if it is L. All successors can infer Asterix's signal perfectly from his decision. If Asterix adopted, then Beatrix, the second individual, should also adopt if her private signal is H; as Beatrix sees it, there have now been two H signals, the one she inferred from Asterix's actions and the one she observed privately. However, if Beatrix's private signal is L, it exactly offsets Asterix's signal H. She is indifferent between adopting and rejecting. We assume, for expositional simplicity, that, as Beatrix is indifferent between the two alternatives, she tosses a coin to decide. (By similar reasoning, if Asterix rejected, then Beatrix should reject if she observes L, and toss a coin if her signal is H.)

The third individual, Cade, faces one of three possible situations: both predecessors adopted (AA), both rejected (RR), or one adopted and the other rejected (AR or RA). In case AA, Cade also adopts. He knows that Asterix observed H and that more likely than not Beatrix observed H too (although she may have seen L and flipped a coin). Thus, even if Cade sees a signal L, he adopts. Consequently, Cade's decision to adopt provides no information to his successors about the desirability of adopting. Cade is therefore in an *information cascade*; his optimal action does not depend on his private information. The uninformativeness of Cade's action means that no further information accumulates. Everyone after Cade faces the same decision and also adopts based only on the observed actions of Asterix and Beatrix. By similar reasoning, RR leads to a cascade of rejection starting with Cade.

In the remaining case where Asterix adopted and Beatrix rejected (or vice versa), Cade knows that Asterix observed H and Beatrix observed L (or vice versa). Thus, Cade's belief based on the actions of the first two individuals is that the $V = 1$ and $V = -1$ are equally likely. He finds himself in a situation identical to that of Asterix, so Cade's

decision is based only on his private signal. Then, the decision problem of the fourth individual, Daisy, is the same as Beatrix's. Asterix's and Beatrix's actions have offset and thus carry no information to Eeyore. And if Cade and Daisy both take the same action – say, adopt – then an adoption cascade starts with Eeyore.

An individual's optimal decision rule is as follows. Let d be the difference between the number of predecessors who adopted and the number who rejected. If $d > 1$, then adopt regardless of private signal. If $d = 1$, then adopt if private signal is H and toss a coin if signal is L. If $d = 0$, then follow private signal. The decisions for $d = -1$ and $d < -1$ are symmetric. The difference between adoptions over rejections evolves randomly, and very quickly hits either the upper barrier of + 2 and triggers an adoption cascade, or the lower barrier of − 2 to trigger a rejection cascade. With virtual certainty, all but the first few individuals end up doing the same thing.

Order of Information, Noise, and Information Externalities

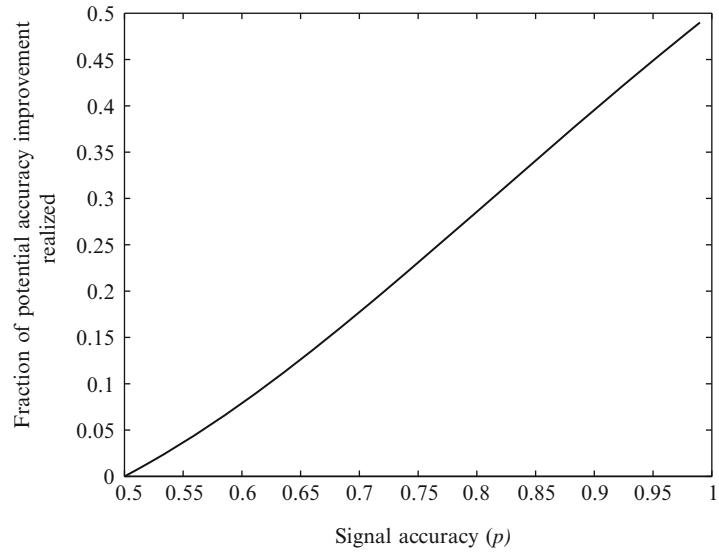
The reason the outcome with observable actions is so different from the observable signals benchmark is that, once a cascade starts, public information stops accumulating. An early preponderance towards adoption or rejection causes all subsequent individuals to ignore their private signals, which thus never join the public pool of knowledge. Nor does the public pool of knowledge have to be very informative to cause individuals to disregard their private signals. As soon as the public pool becomes slightly more informative than the signal of a single individual, individuals defer to the actions of predecessors and a cascade begins.

Furthermore, the type of cascade depends not just on how many H and L signals arrive, but on the order in which they arrive. For example, if signals arrive in the order HHLL. . ., then all individuals adopt, because Cade begins an adoption cascade. If, instead, the same set of signals arrive in the order LLHH. . ., all individuals reject, as Cade begins a rejection cascade. Thus, in the observable actions scenario, whether individuals on the whole adopt or reject is *path dependent*.

A cascade is likely even when private signals are noisy. Specifically, in the above example, let the probability that the signal is correct be $p = 0{:}51$. The probability that an adoption or rejection cascade forms after the first two individuals is close to 75%! (The signal sequences HH – that is, Asterix observes H and Beatrix observes H – and LL cause adoption and rejection cascades respectively, starting with Cade. Similarly, HL and LH each lead to adoption and rejection cascades with probability 0.5 each, if the action chosen by Beatrix after a coin flip is the same as Asterix's. The sum of the probabilities of these events is about 0.75.) After eight players the probability is only 0.004 that the individuals are not in a cascade. (This is the probability $|d| < 2$ for each of individuals 3 through 8.)

Although a cascade starts eventually with probability one, the probability of being in a correct cascade (that is, an adoption cascade when $V = 1$ and a rejection cascade when $V = -1$) is only 0.5133. (The calculation can be found in Bikhchandani et al. 1992.) If individuals do not observe their predecessors' choices (or information), then they would choose an action based only on the private signal; the probability that an individual's choice is correct is 0.51. Thus, the increase in accuracy from observing the actions of predecessors is small. Contrast this with the observable signals scenario, where after many individuals the publicly observed information signals of predecessors are virtually conclusive as to the right action.

More generally, even when individuals have more accurate signals (p is much greater than 0.5), the information contained in a cascade is substantially short of efficient information aggregation. Consider the benchmark observable signals scenario. Individuals far enough out would know the true state almost perfectly. The correctness of these individuals' actions increases from p to 1 due to information revelation. Figure 1 graphs, as a function of the signal accuracy p, the fraction of potential accuracy improvement realized in the observable actions scenario. (The fraction of potential accuracy improvement realized is (Pr[correct



Information Cascades, Fig. 1 Gains from action observability (Note: Fraction of potential accuracy improvement realized $\left(\left[\frac{p(p+1)}{2(1-p+p^2)} - p\right]/(1-p)\right)$ as a function of signal accuracy (p is the probability that the signal is high given that the true value is high))

cascade]-p)(1 -p). From (3) in Bikhchandani et al. 1992, Pr[correct cascade] = $p(p)$ 1) = 2(1 -p) p^2). This fraction increases from 0 for very noisy signals to 0.50 for very informative signals. Thus, in the basic model, at most half of the potential gains are realized.

An individual's private information is useful to others. However, in choosing the optimal action, the individual ignores this benefit: with the onset of a cascade in the observable actions scenario, individuals rationally take uninformative imitative actions. This information externality reduces information aggregation. To see this, consider an alternative benchmark scenario in which (*a*) each individual maximizes a discounted sum of payoffs to all individuals and (*b*) no individual can directly reveal his private information; others learn of his information only through this individual's choice of action. The onset of cascades in this scenario is delayed (compared with the observational actions scenario); information aggregation is efficient subject to the constraint that private information is revealed only through actions.

Fragility

Of course, in reality we do not expect a cascade to last for ever. The arrival of better-informed individuals or the release of new public information can easily dislodge a cascade. Indeed, participants in a cascade know that the cascade is based on information that is only slightly more accurate than the private information of an individual. Thus, a key prediction of the theory is that behaviour in cascades is *fragile* with respect to small shocks. (In some models in which conformity is enforced by the threat of sanctions upon defectors, rare shifts occur when the system crosses a critical value that shifts the outcome from one equilibrium to another; Kuran 1989.)

How robust are the conclusions that cascades are born quickly and idiosyncratically, and shatter easily? When some assumptions in the example are relaxed, is the aggregation of information still inefficient or delayed?

Robustness of the Basic Model

The conclusions of the basic model remain robust along a number of dimensions. We discuss here alternative assumptions about the action space and the signal, which affect the conclusions to some extent.

The Action Space

In the basic model, players make inferences about others' signals from observed choices. When there

are many possible actions, the action choice can convey more information. If the set of actions is continuous and unbounded, then actions fully reveal players' information and cascades do not arise (Lee 1993). (If the action space is a continuous but bounded interval, then when an individual optimally chooses one of the end points of the interval, the value of his signal is not revealed by his action. In consequence, incorrect cascades can form at the end points of the interval.) For example, if a set of firms cannot invest less than zero, they may incorrectly cascade on zero investment.

However, if players are even slightly unsure of the payoff functions of other players, then there is a discontinuous shift to a slower learning process in which information aggregation is inefficient (Vives 1993). In many real-world settings, the action space is bounded or partly discrete: investment projects that have a minimum efficient scale, elections amongst a discrete set of alternatives, a car purchase of a Ford or a Toyota, a takeover decision of whether to bid or not bid for a target firm, and a decision to hire or fire a worker.

The Signal Space

As in the simple two signal example presented above, in settings with a large but discrete set of signal values cascades occur with probability close to one and are sometimes incorrect. In some continuous signal settings cascades do not form (Smith and Sorensen 2000), but an informational externality remains and information aggregation is inefficient. Furthermore, with substantial probability individuals soon follow the behaviour of recent predecessors, and with some probability that action is incorrect. Indeed, with any finite number of individuals, a continuous signal setting is observationally similar to a discrete signal setting that approximates the continuous model. In other words, in a continuous signals setting herds tend to form in which an individual follows the behaviour of his predecessor with high probability, even though this action is not necessarily correct. Thus, the welfare inefficiencies of the discrete cascades model are also present in continuous settings (Chamley 2004, ch. 4).

Observability of Payoffs or Signals

Several papers consider the inefficiency of social learning when there is some degree of observability of payoffs (Caplin and Leahy 1994). Furthermore, even if individuals can observe the payoffs of predecessors, inefficient cascades can form and with positively probability last for ever, because a cascade can lock into an inferior choice before sufficient trials have been performed on the other alternative to persuade later individuals that this alternative is superior (Cao and Hirshleifer 2002). Indeed, if individuals can observe a subset of past signals, such as the past k signals, inefficient cascades can form.

Other Assumptions of Basic Model

When individuals have the freedom to delay their action choice, in equilibrium there is delay, followed by a sudden onset of cascades when an individual commits to an action (Chamley and Gale 1994; Zhang 1997). The existence, idiosyncrasy and fragility of cascades are robust to relaxing other assumptions as well, including allowing for differing information precision, costly information acquisition, and heterogeneous observable tastes (see Bikhchandani et al. 1998, and the references therein). Inefficient cascades still form when individuals have reputational as well as informational motives to herd (Ottaviani and Sorensen 2000). When individuals are imperfectly rational, inefficient cascades still form, but overconfident individuals provide social value when their impetuous choices shatter incorrect cascades (Bernardo and Welch 2001).

Applications

There has been extensive testing of information cascades models in the laboratory. Experiments provide some support for information cascades and observational learning (Anderson and Holt 1997).

Demand for Goods and Securities

The information cascades theory implies not just that consumer purchase decisions will be influenced by others, as occurs, for example, in

automobile purchases in Finland (Grinblatt et al. 2004), but that the source of this influence is informational. In consequence, the cascades approach implies that the incorrect cascades arise in settings in which individuals observe summary statistics of others' behaviour, such as whether one product is outselling another. Golder and Tellis (2004) provide evidence that information cascades play a role in the dynamics of product life cycles. The cascades theory also implies that individuals who are viewed by others as being better informed will be fashion leaders, in the sense that their decisions can trigger immediate cascades. This can explain the effectiveness of a star basketball player's endorsement of a brand of sneakers, but not of his or her endorsement of a brand of beer.

Even without fashion leaders, there are ways for individuals to have disproportionate effects on the onset of information cascades. In a salient 1995 episode, management gurus Michael Treacy and Fred Wiersema secretly purchased 50,000 copies of their business strategy book in order to inflate the sales measures used to construct the *New York Times* best-seller list. Despite mediocre reviews, their book not only made the best-seller list but subsequently sold well enough to continue as a best-seller without further demand intervention by the authors.

The ubiquitous and legitimate marketing method of offering a low initial price may be a successful scheme for introducing an experience good: early adoptions induced by the low price help start a positive cascade. This idea was first analysed by Welch (1992) to explain why initial public offerings of equity are on average severely underpriced by issuing firms. Indeed, a seller may be tempted to cut price secretly for early buyers, so that later buyers will attribute the popularity of the product to high quality rather than low price.

Medicine

Most doctors cannot stay fully abreast of relevant medical advances in their specialties, suggesting that they may select among new treatments based primarily on observation of choices made by other doctors. The cascades approach implies that medical treatments will be characterized by localized conformity and occasional reversals triggered by limited information, and that doctors perceived as having special expertise will have disproportionate influence. It has indeed been claimed that a blind reliance by physicians upon their colleagues' medical decisions commonly leads to surgical fads and even to treatment-caused illnesses (Robin 1984). Many dubious practices seem to have been adopted initially based on weak information (elective hysterectomy, ileal bypass and tonsillectomy), and then later abandoned. A few decades ago, differences in tonsillectomy frequencies in different countries and regions were extreme.

Politics

People learn about others' political beliefs by observing how they vote and from opinion and exit polls. Several studies of political momentum show that early respondents carry disproportionate weight (see Bartels 1988). A possible non-informational explanation is that individuals have a direct preference to conform, but we would expect such an effect to be stronger when an individual is personally exposed to acquaintances with strong views than when the individual observes a polling statistic. Furthermore, polling numbers influence not just preference between candidates, but 'thermometer score' ratings of the perceived quality of candidates. Iowa voters gave an obscure candidate named Jimmy Carter a conspicuous early success in the 1976 US presidential campaign. Many southern states have coordinated their primaries early in the election cycle on the same date ('Super Tuesday') in order to increase their influence on the presidential election. The expanding turnout of protestors in Leipzig in 1989, which triggered the fall of communism in East Germany, has been modelled as an information cascade (Lohmann 1994). More broadly, a recent literature on the social diffusion of ideas emphasizes that individual signals are sometimes not reflected in public discourse, leading to poor information aggregation in public policy decisions (Kuran and Sunstein 1999).

Finance

The decision of individual investors to participate in the stock market and the buying and selling decisions of mutual fund managers are influenced by their peers' decisions (Hong et al. 2005), and there is some indication that herding by mutual funds influences prices (Wermers 1999). The rise in popularity of investment clubs and of day-trading in the 1990s was probably due in part to a self-feeding effect in which individuals learned from the media or word of mouth that many others were day trading. Several theoretical models of securities market trading (Avery and Zemsky 1998) and market crashes (Lee 1998) have been developed which embody either cascade or cascade-like features. Hirshleifer and Teoh (2003) review the theory and evidence of social learning and cascades in finance.

Zoology

Zoologists have documented observational learning, and proposed that information cascades are exhibited in a variety of animal behaviours, including 'false alarm' flights from possible predators, selection of night roosts by birds, and mate-choice copying in various animal species (Giraldeau et al. 2002).

See Also

- ▶ Product Life Cycle
- ▶ Psychology of Social Networks
- ▶ Social Interactions (Empirics)
- ▶ Social Interactions (Theory)
- ▶ Social Norms

Bibliography

Anderson, L., and C. Holt. 1997. Information cascades in the laboratory. *American Economic Review* 87: 847–862.

Avery, C., and P. Zemsky. 1998. Multi-dimensional uncertainty and herd behavior in financial markets. *American Economic Review* 88: 724–748.

Banerjee, A. 1992. A simple model of herd behavior. *Quarterly Journal of Economics* 107: 797–818.

Bartels, L. 1988. *Presidential primaries and the dynamics of public choice*. Princeton: Princeton University Press.

Bernardo, A., and I. Welch. 2001. On the evolution of overconfidence and entrepreneurs. *Journal of Economics and Management Strategy* 10: 301–330.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 1992. A theory of fads, fashion, custom and cultural change as informational cascades. *Journal of Political Economy* 100: 992–1026.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives* 12(3): 151–170.

Cao, H., and D. Hirshleifer. 2002. *Taking the road less traveled: Does conversation eradicate pernicious cascades?* Dice center working paper no. 2002–8. Columbus: Ohio State University.

Caplin, A., and J. Leahy. 1994. Business as usual, market crashes, and wisdom after the fact. *American Economic Review* 84: 548–565.

Chamley, C. 2004. *Rational herds: Economic models of social learning*. Cambridge: Cambridge University Press.

Chamley, C., and D. Gale. 1994. Information revelation and strategic delay in irreversible decisions. *Econometrica* 62: 1065–1085.

Choi, J. 1997. Herd behavior, the penguin effect, and suppression of information diffusion: An analysis of informational externalities and payoff interdependency. *RAND Journal of Economics* 28: 407–425.

Gale, D. 1996. What have we learned from social learning? *European Economic Review* 40: 617–628.

Gibson, R., and J. Hoglund. 1992. Copying and sexual selection. *Trends in Ecological Evolution* 7: 229–232.

Giraldeau, L.-A., T. Valone, and J. Templeton. 2002. Potential disadvantages of using socially acquired information. *Philosophical Transactions of the Royal Society: Biological Sciences* 357: 1559–1566.

Glaeser, E., B. Sacerdote, and J. Scheinkman. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111: 507–548.

Golder, P., and G. Tellis. 2004. Growing, growing, gone: Cascades, diffusion, and turning points in the product life cycle. *Marketing Science* 23: 207–218.

Grinblatt, M., S. Ikäheimo, and M. Keloharju. 2004. *Interpersonal effects in consumption: Evidence from the automobile purchases of neighbors*, Working paper no. 10226. Cambridge, MA: NBER.

Hirshleifer, D., and S. Teoh. 2003. Herding and cascading in capital markets: A review and synthesis. *European Financial Management* 9: 25–66.

Hong, H., J. Kubik, and J. Stein. 2005. Thy neighbor's portfolio: Word-of-mouth effects in the holdings and trades of money managers. *Journal of Finance* 60: 2801–2824.

Kuran, T. 1989. Sparks and prairie fires: A theory of unanticipated political revolution. *Public Choice* 61: 41–74.

Kuran, T., and C. Sunstein. 1999. Availability cascades and risk regulation. *Stanford Law Review* 51: 683–768.

Lee, I. 1993. On the convergence of informational cascades. *Journal of Economic Theory* 61: 396–411.

I

Lee, I. 1998. Market crashes and informational avalanches. *Review of Economic Studies* 65: 741–759.
Lohmann, S. 1994. The dynamics of informational cascades: The Monday demonstrations in Leipzig, East Germany, 1989–91. *World Politics* 47: 42–101.
Machiavelli, N. 1514. *The prince*, ed. Q. Skinner, and R. Price. Cambridge: Cambridge University Press, 1988.
Ottaviani, M., and P. Sorensen. 2000. Herd behavior and investment: Comment. *American Economic Review* 90: 695–704.
Robin, E. 1984. *Matters of life and death: Risks vs. benefits of medical care*. New York: Freeman and Co.
Smith, L., and P. Sorensen. 2000. Pathological outcomes of observational learning. *Econometrica* 68: 371–398.
Taylor, R. 1988. *Medicine out of control: The anatomy of a malignant technology.* Melbourne: Sun Books.
Vives, X. 1993. How fast do rational agents learn? *Review of Economic Studies* 60: 329–347.
Welch, I. 1992. Sequential sales, learning and cascades. *The Journal of Finance* 47: 695–732.
Wermers, R. 1999. Mutual fund herding and the impact on stock prices. *Journal of Finance* 54: 581–622.
Zhang, J. 1997. Strategic delay and the onset of investment cascades. *RAND Journal of Economics* 28: 188–205.

Information Sharing Among Firms

Xavier Vives

Abstract

Firms may have efficiency or strategic incentives to share information about current and past behaviour or intended future conduct. This article examines those incentives and the welfare consequences from the perspective of static oligopoly and monopolistic competition models. It concludes with a review of the available evidence.

Keywords

Antitrust; Bertrand competition; Collusion; Common values; Cournot competition; Disclosure; Disclosure rules; Hard information; Independent values; Information sharing among firms; Misrepresentation; Monopolistic competition; Private values; Product differentiation; Soft information; Trade associations

JEL Classifications

L13

Information sharing (IS) among firms has been a contentious topic in antitrust and has received substantial attention from researchers. Firms may share information about current and past behaviour of, for example, customers, orders and prices, as well as cost and demand conditions. This type of information exchange typically involves hard or verifiable information. Firms may also exchange information about intended future conduct – for example, planned prices, production, new products or capacity expansion. This typically involves soft information. Firms may have incentives to share information for efficiency or strategic reasons. The latter include influencing the behaviour of rivals or sustaining collusion. We will discuss here the results of static models, leaving out dynamic models of collusion and information signalling (see for those models Vives 1999, sects. 8.4, 8.5 and 9.1.5; Kühn and Vives 1995, sect. 8).

Firms may exchange cost or demand information in order to better adapt their output and pricing decisions to uncertainty. From the firm's point of view, the main effects of IS are the increased precision of information to be used by itself and rivals, and the corresponding impact on firms' strategies. In general, increased precision has a positive effect on a firm's expected profits, while the effect of increased precision of rivals and the induced strategy correlation depends on the nature of competition and shocks.

Information exchange is typically modelled as a two-stage game in which firms first unilaterally decide whether to reveal their signals, and then, after receiving those signals and possibly revealing them, compete *à la* Cournot or Bertrand. It is assumed that firms report their signals truthfully if they decide to share information. The workhorse model has quadratic payoffs and normal distributions (or distributions yielding linear conditional expectations) for signals and uncertain parameters such as demand intercepts and marginal costs. The assumptions yield linear equilibria at the second

stage and explicitly computable payoffs. (See Vives 1999, sect. 8.3.1; Kühn and Vives 1995, sects 2–5.) A sample of the literature is Novshek and Sonnenschein (1982), Clarke (1983), Vives (1984), Fried (1984), Gal-Or (1985, 1986), Li (1985), Sakai (1985), Shapiro (1986), Kirby (1988), Sakai and Yamato (1989), Raith (1996), and the extensions in Malueg and Tsutsui (1996, 1998). In the subgame-perfect equilibria of the two-stage game (excepting Bertrand competition with cost uncertainty) unilaterally revealing information is a dominant strategy with independent values, private values (that is, where each firm receives a signal with no noise about its payoff-relevant parameter), or common values with strategic complements. With common value and strategic substitutes, not revealing is a dominant strategy.

If firms are able to enter into industry-wide agreements, the determining factor is whether the information pooling situation increases or reduces expected profits. With the exception of Bertrand competition under cost uncertainty, expected profits with IS are always larger than without, under independent values, private values, and common value and strategic complements. With (for example, Cournot with substitutes), IS yields higher (lower) expected profits for a high (low) degree of product differentiation or steeply (slowly) rising marginal costs. Note that since IS often raises profits under one-shot interaction, IS cannot be taken as prima facie evidence of collusion.

IS agreements are usually mediated by trade associations that typically disclose an aggregate statistic of firms' private signals. Monopolistic competition, where no firm has a significant impact on aggregate market outcomes, is suitable for examining the role of such associations' disclosure rules. A firm first decides whether or not to join the association and reveal its private information. Under non- exclusionary disclosure, information is made available to everyone in the market; under exclusionary disclosure, it is provided to members only. Obviously, with a non-exclusionary disclosure rule, IS will not ensue if the sharing is costly (by not joining, a firm, being negligible in terms of aggregate market impact, can free ride and obtain market information costlessly, with no effect on market aggregates). With an exclusionary disclosure rule, IS may occur if the membership fee is not too high (see Vives 1990).

The impact of IS on consumer surplus and total surplus depends on the type of competition and uncertainty, and on the number of firms. Three effects operate: output adjustment to information, output uniformity across varieties (given consumer preference for variety), and selection among firms of different efficiencies. IS may allow firms to better adjust to demand and/or costs shocks (output adjustment effect). This will tend to improve welfare except if the firm is a price setter and demand is uncertain. In this case, more information will give the firm greater scope to extract consumer surplus – an insight already valid for a monopolist. In monopolistic competition, where variety must be taken into account, IS tends to make the outputs of varieties more similar with common value uncertainty and less so with private value uncertainty, thus increasing (decreasing) expected total surplus under demand uncertainty and Cournot (Bertrand) competition (Vives 1990).

Analysis of the oligopoly case is complex, but several generalizations hold. Under demand uncertainty and Cournot competition, IS increases expected total surplus (ETS); under demand uncertainty and Bertrand competition, it decreases consumer surplus (as well as ETS, under monopolistic competition). With common values, IS always increases ETS, except under price competition, when goods are poor substitutes and/or there are many firms. (See Kühn and Vives 1995, sect. 5.2; Vives 1999, sect. 8.3.3.) There are potentially large efficiency benefits from information exchange. For example, the production rationalization effect of cost information exchange under Cournot can be very large and is of a larger order of magnitude than the market power effect (Vives 2002).

What happens when there is no trade association to provide a mechanism to share information truthfully? Assume private cost information that is exchangeable only at an interim stage, once each firm learns its own cost but does not know its rivals'. In this case, if information is not verifiable

and there are no other signalling possibilities, information revelation is impossible, since all firms would like to be perceived as being low-cost. With verifiable information, full revelation ensues if disclosure is costless and it is known whether firms have information (Okuno- Fujiwara et al. 1990; Van Zandt and Vives 2006). The lowest- cost firm will reveal its type and then all other types will unravel. Information could also be revealed through costly signalling in the form of wasteful advertising (for example, Ziv 1993), or via dynamic competition in which production levels are observable (Mailath 1989) or with sales reports (Jin 1994). In the latter case, sharing sales reports eliminates the incentive to misrepresent and changes the consequences of IS. If it is possible to verify information but not whether the firm is informed, then the unravelling result need not hold, and firms can selectively disclose acquired information (Jansen 2005).

Evidence on the effect of IS among firms is scant. Genesove and Mullin (1999) study information exchange in the Sugar Institute and find no misreporting, but some information withholding, suggesting that information can be verified. Doyle and Snyder (1999) study production plans announcements in the trade press in the automobile industry and find that a firm's announcement affects competitors' responses. Announcements of increased production are met by upward adjustments in production, which they interpret as consistent with announcements signalling a common demand parameter. Christensen and Caves (1997) study capacity announcements in the pulp and paper industry and find that unexpected announcements by rivals promote project abandonment in sub-industries with low concentration levels (and the opposite in concentrated sub-industries); they compare these results with IS models of cost information. Armantier and Richard (2003) examine exchange of cost information in the multi-market context of the airline industry. The authors account for entry decisions in a Cournot setting with complementary goods across markets, and simulate a hypothetical agreement to share cost information by American Airlines and United Airlines at Chicago O'Hare airport. They find that IS would improve airline profitability and moderately harm consumers (although, theoretically, cost IS need not necessarily hurt consumers in such a situation). The experimental results in Cason (1994) suggest that pricing behaviour is influenced by IS decisions. Ackert et al. (2000) find that in a Cournot game with cost uncertainty, where it cannot be verified whether a firm has received information, when a firm receives information about industry-wide cost unfavourable information is disclosed but favourable information is withheld. Contrary to theory, when information is about a cost-specific shock, disclosure is not affected by the favourableness of information.

See Also

► Cartels

Bibliography

Ackert, L., B. Church, and M. Sankar. 2000. Voluntary disclosure under imperfect competition: Experimental evidence. *International Journal of Industrial Organization* 18: 81–105.

Armantier, O., and O. Richard. 2003. Exchanges of cost information in the airline industry. *RAND Journal of Economics* 34: 461–477.

Cason, T. 1994. The impact of information sharing opportunities on market outcomes: An experimental study. *Southern Economic Journal* 61: 18–39.

Christensen, L., and R. Caves. 1997. Cheap talk and investment rivalry in the pulp and paper industry. *Journal of Industrial Economics* 45: 47–73.

Clarke, R. 1983. Collusion and the incentives for information sharing. *Bell Journal of Economics* 14: 383–394.

Doyle, P., and C. Snyder. 1999. Information sharing and competition in the motor vehicle industry. *Journal of Political Economy* 107: 1326–1364.

Fried, D. 1984. Incentives for information production and disclosure in a duopolistic environment. *Quarterly Journal of Economics* 99: 367–381.

Gal-Or, E. 1985. Information sharing in oligopoly. *Econometrica* 53: 329–343.

Gal-Or, E. 1986. Information transmission-Cournot and Bertrand equilibria. *Review of Economic Studies* 53: 85–92.

Genesove, D., and W. Mullin. 1999. The Sugar Institute learns to organize information exchange. In *Learning by doing in markets, firms and countries*, ed. N. Lamoreaux, D. Raff, and P. Temin. Chicago: University of Chicago Press for the NBER.

Jansen, J. 2005. *Information acquisition and strategic disclosure in oligopoly*. Berlin: Mimeo Social Science Research Center Berlin (WZB) and Humboldt University.
Jin, Y. 1994. Information sharing through sales report. *Journal of Industrial Economics* 42: 323–333.
Kirby, A. 1988. Trade associations as information exchange mechanisms. *RAND Journal of Economics* 19: 138–146.
Kühn, K., and X. Vives. 1995. *Information exchanges among firms and their impact on competition*. Luxembourg: Office for Official Publications of the European Communities.
Li, L. 1985. Cournot oligopoly with information sharing. *RAND Journal of Economics* 16: 521–536.
Mailath, G. 1989. Simultaneous signalling in an oligopoly model. *Quarterly Journal of Economics* 104: 417–427.
Malueg, D., and S. Tsutsui. 1996. Duopoly information exchange: The case of unknown slope. *International Journal of Industrial Organization* 14: 119–136.
Malueg, D., and S. Tsutsui. 1998. Distributional assumptions in the theory of oligopoly information exchange. *International Journal of Industrial Organization* 16: 785–797.
Novshek, W., and H. Sonnenschein. 1982. Fulfilled expectations Cournot duopoly with information acquisition and release. *Bell Journal of Economics* 13: 214–218.
Okuno-Fujiwara, M., A. Postlewaite, and K. Suzumura. 1990. Strategic information revelation. *Review of Economic Studies* 57: 25–47.
Raith, M. 1996. A general model of information sharing in oligopoly. *Journal of Economic Theory* 71: 260–288.
Sakai, Y. 1985. The value of information in a simple duopoly model. *Journal of Economic Theory* 36: 36–54.
Sakai, Y., and T. Yamato. 1989. Oligopoly, information and welfare. *Journal of Economics (Zeitschrift für Nationalökonomie)* 49: 3–24.
Shapiro, C. 1986. Exchange of cost information in oligopoly. *Review of Economic Studies* 53: 433–446.
Van Zandt, T., and X. Vives. 2006. Monotone equilibria in Bayesian games of strategic complementarities. *Journal of Economic Theory* (forthcoming).
Vives, X. 1984. Duopoly information equilibrium: Cournot and Bertrand. *Journal of Economic Theory* 34: 71–94.
Vives, X. 1990. Trade association, disclosure rules, incentives to share information and welfare. *RAND Journal of Economics* 22: 446–453.
Vives, X. 1999. *Oligopoly pricing: Old ideas and new tools*. Boston: MIT Press.
Vives, X. 2002. Private information, strategic behavior, and efficiency in Cournot markets. *RAND Journal of Economics* 33: 361–376.
Ziv, A. 1993. Information sharing in oligopoly: The truth-telling problem. *RAND Journal of Economics* 24: 455–465.

Information Technology and the World Economy

Dale W. Jorgenson and Khuong Vu

Abstract

This article analyses the impact of investment in information technology (IT) on the recent resurgence of world economic growth. We describe the growth of the world economy, seven regions, and 14 major economies during the period 1989–2004. We allocate the growth of world output between input growth and productivity and find, surprisingly, that input growth greatly predominates. Moreover, differences in per capita output levels are explained by differences in per capita input rather than variations in productivity. The contributions of IT investment have increased in all regions, but especially in industrialized economies and Developing Asia.

Keywords

Asian miracle; Growth accounting; Human capital; Information technology and the world economy; Input growth; Productivity growth

JEL Classifications

O4

Introduction

This article analyses the impact of investment in information technology (IT) equipment and software on the recent resurgence in world economic growth. The crucial role of IT investment in the growth of the US economy has been thoroughly documented and widely discussed. (See Jorgenson and Stiroh 2000, and Oliner and Sichel 2000. The growth accounting methodology employed in this literature is discussed by Jorgenson et al. 2005, and summarized by

Jorgenson, 2005.) Jorgenson (2001) has shown that the remarkable behaviour of IT prices is the key to understanding the resurgence of American economic growth. This behaviour can be traced to developments in semiconductor technology that are widely understood by technologists and economists.

Jorgenson (2003) has shown that the growth of IT investment jumped to double-digit levels after 1995 in all the G7 economies – Canada, France, Germany, Italy, Japan, and the United Kingdom, as well as the United States. (Ahmad et al. 2004, have analysed the impact of IT investment in OECD countries. van Ark et al. 2003; 2005), and Francesco Daveri 2002, have presented comparisons among European economies.) These economies account for nearly half of world output and a much larger share of world IT investment. The surge of IT investment after 1995 resulted from a sharp acceleration in the rate of decline of prices of IT equipment and software. Jorgenson (2001) has traced this to a drastic shortening of the product cycle for semiconductors from three years to two years, beginning in 1995.

In section "World Economic Growth, 1989–2003" we describe the growth of the world economy, seven economic regions, and 14 major economies given in Table 1 during the period 1989–2003. (We include 110 economies with more than one million in population and a complete set of national accounts for the period 1989–2003 from Penn World Table, 2002, and World Bank Development Indicators Online, 2004. These economies account for more that 96 per cent of world output.) The world economy is divided among the G7 and Non-G7 industrialized economies, Developing Asia, Latin America, Eastern Europe and the former Soviet Union, North Africa and the Middle East, and sub-Saharan Africa. The 14 major economies include the G7 economies listed above and the developing and transition economies of Brazil, China, India, Indonesia, Mexico, Russia, and South Korea.

We have sub-divided the period in 1995 in order to focus on the response of IT investment to the accelerated decline in IT prices. As shown in Table 1, world economic growth has undergone a powerful revival since 1995. The per capita growth rate jumped nearly a full percentage point from 2.50 per cent during 1989–95 to 3.45 per cent in 1995–2003. We can underline the significance of this difference by pointing out that per capita growth of 3.45 per cent doubles world output per capita in a little over two decades, while slower growth of 2.50 per cent doubles per capita output in slightly less than three decades.

In section "Sources of World Economic Growth" we allocate the growth of world output between input growth and productivity. Our most astonishing finding is that input growth greatly predominated! Productivity growth contributed only one-fifth of the total during 1989–95, while input growth accounted for almost four-fifths. Similarly, input growth contributed more than 70 per cent of growth after 1995, while productivity accounted for less than 30 per cent. The only important departure from this worldwide trend is the Asian miracle before 1995, when the rate of economic growth in Developing Asia far outstripped the rest of the world and productivity growth predominated.

In section "Sources of World Economic Growth" we distribute the growth of input per capita between investments in tangible assets, especially IT equipment and software, and investments in human capital. The world economy, all seven regions, and the 14 major economies, except Indonesia and Mexico, experienced a surge in investment in IT after 1995. The soaring level of US IT investment after 1995 was paralleled by jumps in IT investment throughout the industrialized world. The contributions of IT investment in Developing Asia, Latin America, Eastern Europe, North Africa and the Middle East, and sub-Saharan Africa more than doubled after 1995, beginning from much lower levels. By far the most dramatic increase took place in Developing Asia.

In section "World Output, Input and Productivity" we present levels of output per capita, input per capita and productivity for the world economy, the seven economic regions, and the 14 major economies. We find that differences in per capita output levels are primarily explained by

differences in per capita input, rather than variations in productivity. Taking US output per capita in 2000 as 100.0, world output per capita was a relatively modest 23.9 in 2003. If we use similar scales for input and productivity, world input per capita in 2003 was a substantial 42.4 and world productivity a robust 56.3. Section "Summary and Conclusions" concludes the paper.

World Economic Growth, 1989–2003

In order to set the stage for analysing the impact of IT investment on the growth of the world economy, we first consider the shares of world product and growth for each of the seven regions and the 14 major economies presented in Table 1. Following Jorgenson (2001), we have chosen GDP as a

Information Technology and the World Economy, Table 1 The world economy: shares in size and growth by group, region, and major economies. The measures for groups and the world are averages weighted by GDP (in PPP$) share

	Period 1989–1995			Period 1995–2003		
		Average share			Average share	
Group/region	GDP growth	GDP	Growth	GDP growth	GDP	Growth
World (110 economies)	**2.50**	**100.00**	**100.00**	**3.45**	**100.00**	**100.00**
G7 (7 economies)	2.18	47.44	41.33	2.56	45.26	33.62
Developing Asia (16)	7.35	20.76	61.13	5.62	26.05	42.56
Non-G7 (15)	2.03	8.38	6.77	3.01	8.13	7.10
Latin America (19)	3.06	8.35	10.20	2.11	8.07	4.94
Eastern Europe (14)	− 7.05	9.32	−26.76	2.87	6.57	5.47
Sub-Saharan Africa (28)	1.21	2.13	1.03	2.88	2.01	1.68
N. Africa and Middle East (11)	4.36	3.61	6.29	4.08	3.91	4.64

	Period 1989–1995					Period 1995–2003				
	GDP	Avg. GDP share		Growth share		GDP	Avg. GDP share		Growth share	
Economy	growth	Group	World	Group	World	growth	Group	World	Group	World
Seven world major economies (G7)										
Canada	1.39	4.90	2.32	3.12	1.29	2.51	4.78	2.17	4.69	1.58
France	1.30	7.10	3.37	4.23	1.75	1.92	6.76	3.06	5.05	1.70
Germany	2.34	10.80	5.12	11.58	4.79	0.86	10.20	4.63	3.41	1.15
Italy	1.52	7.42	3.52	5.17	2.14	1.48	6.99	3.17	4.05	1.36
Japan	2.56	16.23	7.70	19.03	7.88	1.39	15.73	7.13	8.54	2.88
United	1.62	7.45	3.54	5.53	2.29	2.55	7.37	3.34	7.32	2.46
Kingdom										
United States	2.43	46.11	21.87	51.34	21.26	3.56	48.16	21.76	66.92	22.46
All G7	**2.18**	**100.0**	**47.4**	**100.00**	**41.4**	**2.56**	**100.0**	**45.3**	**100.0**	**33.6**
	Period 1989–1995					Period 1995–2003				
	GDP	Avg. GDP share		Growth share		GDP	Avg. GDP share		Growth share	
Economy	growth	Group	World	Group	World	growth	Group	World	Group	World
Seven major developing and transition economies (GD7)										
Brazil	1.97	12.1	3.16	6.89	2.48	1.94	10.16	2.93	3.8	1.65
China	9.94	29.26	7.64	84.23	30.36	7.13	37.79	10.91	51.99	22.55
India	5.03	18.98	4.95	27.65	9.97	6.15	20.69	5.97	24.54	10.65
Indonesia	6.82	7.12	1.86	14.07	5.07	2.41	6.98	2.02	3.25	1.41
Mexico	2.19	7.48	1.95	4.74	1.71	3.56	6.74	1.95	4.63	2.01
Russian Federation	− 8.44	19.92	5.2	-48.71	- 17.56	3.18	12.17	3.52	7.46	3.24
South Korea	7.48	5.14	1.34	11.13	4.01	4.09	5.47	1.58	4.32	1.87
All GD7	**3.45**	**100**	**26.1**	**100**	**36**	**5.18**	**100**	**28.9**	**100**	**43.4**

measure of output. We employ the Penn World Table, generated by Heston et al. (2002), as the primary data source on GDP and purchasing power parities for economies outside the G7 and the European Union, as it existed prior to enlargement in May 2004. (Maddison 2001, provides estimates of national product and population for 134 countries for varying periods from 1820 to 1998 in his magisterial volume, *The World Economy: A Millennial Perspective*.)

We have revised and updated the US data presented by Jorgenson (2001) through 2003. Comparable data for Canada have been constructed by Statistics Canada (see Baldwin and Harchaoui 2003). Data for France, Germany, Italy, and the UK and the economies of the European Union before enlargement have been developed for the European Commission by Bart van Ark et al. (2003). Finally, data for Japan have been assembled by Jorgenson and Kazuyuki Motohashi (2005) for the Research Institute on Economy, Trade, and Industry. We have linked these data by means of the OECD's purchasing power parities for 1999 (OECD 2002).

The G7 economies accounted for slightly under half of world product from 1989 to 2003. The per capita growth rates of these economies – 2.18 per cent before 1995 and 2.56 per cent afterward – were considerably below world growth rates. The growth acceleration of 0.60 per cent for the G7 economies lagged behind the jump in world economic growth. The G7 shares in world growth were 41.3 per cent during 1989–95 and 33.6 per cent in 1995–2003, well below the G7 shares in world product of 47.4 per cent and 45.3 per cent, respectively.

During 1995–2003 the United States accounted for 21.8 per cent of world product and 48.2 per cent of G7 output. After 1995 Japan fell from its ranking as the world's second largest economy to third largest after China. Germany dropped from fourth place before 1995, following the United States, China and Japan, to fifth place during 1995–2003, ranking behind India as well. Japan remained the second largest of the G7 economies, while Germany retained its position as the leading European economy. France, Italy and the UK were similar in size, but less than half the size of Japan. Canada was the smallest of the G7 economies.

The US growth rate jumped from 2.43 per cent during 1989–95 to 3.56 per cent in 1995–2003. The period 1995–2003 included the shallow US recession of 2001 and the ensuing recovery, as well as the IT-generated investment boom of the last half of the 1990s. The United States accounted for more than half of G7 growth before 1995 and more than two-thirds afterward. The US share in world growth fell below its share in world product before 1995, but rose above the US product share after 1995. By contrast Japan's share in world economic growth before 1995 exceeded its share in world product, but fell short of the product share after 1995. The remaining G7 economies had lower shares of world growth than world product before and after 1995.

The 16 economies of Developing Asia generated slightly more than a fifth of world output before 1995 and more than a quarter afterward. The burgeoning economies of China and India accounted for more than 60 per cent of Asian output in both periods. (Our data for China are taken from the Penn World Table, 2002. Alwyn Young 2003, presents persuasive evidence that the official estimates given, for example, by the World Development Indicators, 2004, exaggerate the growth of output and productivity in China.) The economies of Developing Asia grew at 7.35 per cent before 1995 and 5.62 per cent afterward. These economies were responsible for an astounding 61 per cent of world growth during 1989–95! Slightly less than half of this took place in China, while a little less than a third occurred in India. Developing Asia's share in world growth declined to 43 per cent during 1995–2003, remaining well above the region's share of 26.1 per cent of world product. China accounted for more than half of this growth and India about a quarter.

The 15 Non-G7 industrialized economies generated more than eight per cent of world output during 1989–2003. These economies were responsible for lower shares in world growth than world product before and after 1995. Prior to the fall of the Berlin Wall and the collapse of the Soviet Union, the 14 economies of Eastern

Europe and the former Soviet Union were larger in size than the Non-G7, generating 9.3 per cent of world product. All of the economies of Eastern Europe experienced a decline in output during 1989–95. Collectively, these economies subtracted 26.8 per cent from world growth during 1989–95, dragging their share of world product down to 6.6 per cent. During 1989–1995 Russia's economy was comparable in size to Germany's, but from 1995 to 2003 the Russian economy was only slightly larger than the UK economy.

During 1989–95 the ten per cent share of the Latin American economies in world growth exceeded their eight-and-a-half percent share in world product. After 1995 these economies had a substantially smaller six per cent share in world growth, while retaining close to an eight-and-a-half share in world product with Brazil and Mexico responsible for more than 60 per cent of this. Brazil's share in world growth was below its three per cent share in world product before and after 1995, while Mexico's growth was lower than its product share before 1995 and higher afterward.

The 11 economies of North Africa and the Middle East, taken together, were comparable in size to France, Italy, or the UK, while the 30 economies of sub-Saharan Africa, as a group, ranked with Canada. The economies of North Africa and the Middle East had a share in world growth of 6.3 per cent during 1989–95, well above their 3.6 per cent share in world product. After 1995 their share in world growth fell to 4.6 per cent, still above the share in world product of 3.9 per cent. Growth in the economies of sub-Saharan Africa lagged behind their shares in world product during both periods.

Sources of World Economic Growth

We next allocate the sources of world economic growth during 1989–2003 between the contributions of capital and labour inputs and the growth of productivity. We find that productivity, frequently touted as the primary engine of economic growth, accounted for only 20–30 per cent of world growth. Nearly half of this growth can be attributed to the accumulation and deployment of capital and another a quarter to a third to the more effective use of labour. Our second objective is to explore the determinants of the growth of capital and labour inputs, emphasizing the role of investment in information technology equipment and software and the importance of investment in human capital.

We have derived estimates of capital input and property income from national accounting data for the G7 economies. We have constructed estimates of hours worked and labour compensation from labour force surveys for each of these economies. We measure the contribution of labour inputs, classified by age, sex, educational attainment, and employment status, by weighting the growth rate of each type of labour input by its share in the value of output. Finally, we employ purchasing power parities for capital and labour inputs constructed by Jorgenson (2003). (Purchasing power parities for inputs follow the methodology described in detail by Jorgenson and Yip 2000.)

We have extended these estimates of capital and labour inputs to the 103 Non-G7 countries using data sources and methods described in section "Methods and Data Sources". (We employ data on educational attainment from Barro and Lee 2001, and governance indicators constructed by Kaufmann et al. 2004, for the World Bank; for further details, see section "Methods and Data Sources".)

We have distinguished investments in information technology equipment and software from investments in other assets for all 110 economies in our study. We have derived estimates of IT investment from national accounting data for the G7 economies and those of the European Union before enlargement. We measure the contribution of IT investment to economic growth by weighting the growth rate of IT capital inputs by the shares of these inputs in the value of output. Similarly, the contribution of Non-IT investment is a share-weighted growth rate of Non-IT capital inputs. The contribution of capital input is the sum of these two components.

We have revised and updated the US data presented by Jorgenson (2001) on investment in information technology and equipment. (US data

on investment in IT equipment and software, provided by the Bureau of Economic Analysis, BEA, are the most comprehensive and detailed. The BEA data are described by Grimm et al. 2005.) Data on IT investment for Canada have been have been constructed by Statistics Canada (Baldwin and Harchaoui 2003). Data for the countries of the European Union have been developed for the European Commission by van Ark et al. (2003). Finally, data for Japan have been assembled by Jorgenson and Motohashi (2005). We have relied on the WITSA *Digital Planet Report* (2002/2004) as the starting point for estimates of IT investment for the remaining economies. (WITSA stands for the World Information Technology and Services Alliance. Other important sources of data include the International Telecommunication Union, ITU, telecommunications indicators, the UNDP *Human Development* reports, and the Business Software Alliance 2003. Additional details are given in section "Methods and Data Sources".)

We have divided labour input growth between the growth of hours worked and labour quality, where quality is defined as the ratio of labour input to hours worked. This reflects changes in the composition of labour input, for example, through increases in the education and experience of the labour force. The contribution of labour input is the rate of growth of this input, weighted by the share of labour in the value of output. Finally, productivity growth is the difference between the rate of growth of output and the contributions of capital and labour inputs.

The contribution of capital input to world economic growth before 1995 was 1.18 per cent, a little more than 47 per cent of the growth rate of 2.50 per cent. Labour input contributed 0.79 per cent or slightly less than 32 per cent, while productivity growth of 0.53 per cent or just over 21 per cent. After 1995 the contribution of capital input climbed to 1.56 per cent, around 45 per cent of output growth, while the contribution of labour input rose to 0.89 per cent, around 26 per cent. Productivity increased to 0.99 per cent or nearly 29 per cent of growth. We arrive at the astonishing conclusion that the contributions of capital and labour inputs greatly predominated over productivity as sources of world economic growth before and after 1995!

We have divided the contribution of capital input to world economic growth between IT equipment and software and Non-IT capital input. The contribution of IT almost doubled after 1995, less than a quarter to more than a third of the contribution of capital input. However, Non-IT was more important before and after 1995. We have divided the contribution of labour input between hours worked and labour quality. Hours rose from 0.39 per cent before 1995 to 0.62 per cent after 1995, while labour quality declined from 0.40 per cent to 0.27 per cent. Labour quality and hours worked were almost equal in importance before 1995, but hours worked became the major source of labour input growth after 1995.

The acceleration in the world growth rate after 1995 was 0.95 per cent, almost a full percentage point. The contribution of capital input explained 0.38 per cent of this increase, while the productivity accounted for 0.46 per cent. Labour input contributed a relatively modest 0.10 per cent. The jump in IT investment of 0.26 per cent was most important source of the increase in capital input. This can be traced to the stepped-up rate of decline of IT prices after 1995 analysed by Jorgenson (2001). The substantial increase of 0.23 per cent in the contribution of hours worked was the most important component of labour input growth.

Table 2 presents the contribution of capital input to economic growth for the G7 economies, divided between IT and Non-IT. Capital input was the most important source of growth before and after 1995. The contribution of capital input before 1995 was 1.28 or almost three-fifths of the G7 growth rate of 2.18 per cent, while the contribution of 1.43 per cent after 1995 was 55 per cent of the higher growth rate of 2.56 per cent. Labour input growth contributed 0.49 per cent before 1995 and 0.46 per cent afterward, about 22 per cent and 18 per cent of growth, respectively. Productivity accounted for 0.42 per cent before 1995 and 0.67 per cent after 1995 or less than a fifth and slightly more than a quarter of G7 growth, respectively.

The powerful surge of IT investment in the United States after 1995 is mirrored in jumps in

Information Technology and the World Economy, Table 2 Sources of output growth: 1995–2003 vs. 1989–1995. The measures for groups and the world are averages weighted by GDP (in PPP$) share GDP Sources of growth (% points per GDP Sources of growth (% points per Growth annum) Growth annum)

Economy	Period 1989–1995						Period 1995–2003					
	GDP Growth	Sources of growth (% points per annum)					GDP Growth	Sources of growth (% points per annum)				
		Capital		Labour		TFP		Capital		Labour		TFP
		ICT	Non-ICT	Hours	Quality			ICT	Non-ICT	Hours	Quality	
World (110 economies)	**2.50**	**0.27**	**0.91**	**0.39**	**0.40**	**0.53**	**3.45**	**0.53**	**1.03**	**0.62**	**0.27**	**0.99**
G7 (7 economies)	2.18	0.38	0.90	0.07	0.42	0.42	2.56	0.69	0.74	0.28	0.18	0.67
Developing Asia (16)	7.35	0.15	1.73	1.19	0.42	3.86	5.62	0.43	2.27	0.81	0.38	1.72
Non-G7 (15)	2.03	0.32	0.68	0.21	0.21	0.61	3.01	0.49	0.77	1.06	0.20	0.49
Latin America (19)	3.06	0.16	0.58	1.20	0.37	0.75	2.11	0.39	0.61	1.10	0.34	−0.32
Eastern Europe (14)	− 7.05	0.10	−0.15	−0.86	0.36	−6.50	2.87	0.23	− 0.81	0.01	0.39	3.06
Sub-Saharan Africa (28)	1.21	0.3	0.24	1.66	0.56	−1.39	2.88	0.29	0.68	1.18	0.42	0.32
N. Africa and Middle East (11)	4.36	0.15	0.72	1.43	0.56	1.50	4.08	0.40	0.88	2.02	0.49	0.30
Seven world major economies (G7)												
Canada	1.39	0.49	0.27	0.07	0.55	0.01	2.51	0.65	0.61	0.68	0.16	0.42
France	1.30	0.19	0.93	−0.17	0.61	−0.26	1.92	0.36	0.75	0.22	0.07	0.52
Germany	2.34	0.26	1.05	−0.42	0.33	1.12	0.86	0.40	0.50	−0.24	0.09	0.11
Italy	1.52	0.26	0.86	−0.35	0.38	0.37	1.48	0.46	0.96	0.61	0.27	−0.82
Japan	2.56	0.31	1.16	−0.39	0.54	0.94	1.39	0.56	0.26	−0.32	0.22	0.67
United Kingdom	1.62	0.27	1.69	−0.73	0.49	−0.10	2.55	0.65	0.19	0.38	0.26	1.07
United States	2.43	0.49	0.71	0.57	0.36	0.31	3.56	0.88	1.01	0.50	0.17	0.99
All G7	**2.18**	**0.38**	**0.90**	**0.07**	**0.42**	**0.42**	**2.56**	**0.69**	**0.74**	**0.28**	**0.18**	**0.67**
Seven major developing and transition economies (GD7)												
Brazil	1.97	0.09	0.29	0.99	0.39	0.20	1.94	0.46	0.24	0.67	0.37	0.21
China	9.94	0.17	2.12	0.87	0.45	6.33	7.13	0.63	3.17	0.45	0.39	2.49
India	5.03	0.09	1.18	1.27	0.43	2.06	6.15	0.26	1.77	1.22	0.41	2.49
Indonesia	6.82	0.10	1.62	1.64	0.43	3.04	2.41	0.09	1.47	0.91	0.41	−0.47
Mexico	2.19	0.24	0.95	1.48	0.38	−0.87	3.56	0.23	1.11	1.76	0.31	0.14
Russian Federation	− 8.44	0.07	−0.07	−1.02	0.37	−7.79	3.18	0.10	−1.30	0.21	0.44	3.73
South Korea	7.48	0.29	2.31	1.45	0.31	3.13	4.09	0.46	1.67	0.86	0.26	0.85
All GD7	**3.45**	**0.13**	**1.17**	**0.72**	**0.41**	**1.03**	**5.18**	**0.40**	**1.70**	**0.74**	**0.39**	**1.96**

the growth rates of IT capital through the G7. The contribution of IT capital input for the G7 increased from 0.38 during the period 1989–95 to 0.69 per cent during 1995–2003, rising from 30 per cent of the contribution of capital input to more than 48 per cent. The contribution of Non-IT capital input predominated in both periods, but receded slightly from 0.90 per cent before 1995 to 0.74 per cent afterward. This reflected more rapid substitution of IT capital input for Non-IT capital input in response to swiftly declining prices of IT equipment and software after 1995.

The modest acceleration of 0.38 per cent in G7 output growth after 1995 was powered by investment in IT equipment and software, accounting for 0.31 per cent, while the contribution of Non-IT investment slipped by 0.16 per cent. Before 1995 the contribution of labour quality of 0.42 per cent accounted for more than 80 per cent of the contribution of G7 labour input, while the contribution of hours worked of 0.28 per cent explained more than 60 per cent after 1995. The rising contribution of hours worked was offset by the declining contribution of labour quality, while productivity growth rose by 0.25 per cent.

In Developing Asia the contribution of capital input increased from 1.88 per cent before 1995 to 2.70 per cent after 1995, while the contribution of labour input fell from 1.61 per cent to 1.19 per cent. These opposing trends had a slightly positive impact on growth. The significant slowdown in the Asian growth rate from 7.35 per cent to 5.62 per cent can be traced entirely to a sharp decline in productivity growth from 3.86 to 1.72 per cent. Productivity explained slightly over half of Asian growth before 1995, but only 30 per cent after 1995.

The first half of the 1990s was a continuation of the Asian Miracle, analysed by Krugman (1994), Lau (1999), and Young (1995). This period was dominated by the spectacular rise of China and India and the continuing emergence of the Gang of Four – Hong Kong, Singapore, South Korea, and Taiwan. However, all the Asian economies had growth rates considerably in excess of the world average of 2.50 per cent. The second half of the 1990s was dominated by the Asian financial crisis but, surprisingly, conforms much more closely to the 'Krugman thesis' attributing Asian growth to input growth rather than productivity.

The Krugman thesis was originally propounded to distinguish the Asian Miracle from growth in industrialized countries. According to this thesis, Asian growth was differentiated by high growth rates and a great predominance of inputs over productivity as the sources of high growth. In fact, productivity growth exceeded the growth of input during the Asian Miracle of the early 1990s! Moreover, growth in the world economy and the G7 economies was dominated by growth of capital and labour inputs before and after 1995. Productivity growth played a subordinate role and fell considerably short of the contributions of capital and labour inputs to world and G7 growth.

Developing Asia experienced a potent surge in investment in IT equipment and software after 1995. The contribution of IT investment more than doubled from 0.15 percent to 0.43 per cent, explaining less than eight per cent of the contribution of capital input before 1995, but almost 16 per cent afterward. The rush in IT investment was particularly powerful in China, rising from 0.17 per cent before 1995 to 0.63 per cent afterward. India fell substantially behind China, but outperformed the region as a whole, increasing the contribution of IT investment from 0.09 to 0.26 per cent.

Indonesia was the only major economy to experience a decline in the contribution of both IT and Non-IT investment after 1995. South Korea's IT investment increased from 0.29 before 1995 to 0.46 per cent afterward, while Non-IT investment dropped as a consequence of the Asian financial crisis. The contribution of Non-IT investment in Asia greatly predominated in both periods and also accounted for most of the increase in the contribution of capital input after 1995. The contributions of hours worked and labour quality declined after 1995 with hours worked dominating in both periods.

Economic growth in the 15 Non-G7 industrialized economies accelerated much more sharply than G7 growth after 1995. The contribution of labour input slightly predominated over capital

input before and after 1995. The contribution of labour input was 0.81 per cent before 1995, accounting for about 40 per cent of Non-G7 growth, and 1.26 after 1995, explaining 39 per cent of growth. The corresponding contributions of capital input were 0.75 per cent and 1.12 per cent, explaining 37 and 34 per cent of Non-G7 growth, respectively. Non-G7 productivity also rose from 0.47 before 1995 to 0.89 percent afterward; however, productivity accounted for only 23 and 27 per cent of growth in these two periods.

The impact of investment in IT equipment and software in the Non-G7 economies doubled after 1995, rising from 0.22 per cent to 0.44 per cent or from 29 per cent of the contribution of Non-G7 capital input to 39 per cent. This provided a substantial impetus to the acceleration in Non-G7 growth of 1.25 per cent. Non-IT investment explained another 0.14 per cent of the growth acceleration. However, the increased contribution of hours worked of 0.49 per cent and improved productivity growth of 0.42 per cent predominated.

The collapse of economic growth in Eastern Europe and the former Soviet Union before 1995 can be attributed almost entirely to a steep decline in productivity during the transition from socialism. This was followed by a modest revival in both growth and productivity after 1995, bringing many of the transition economies close to levels of output per capita that prevailed in 1989. The contribution of capital input declined both before and after 1995, even as the contribution of IT investment jumped from 0.09 to 0.26 per cent. Hour worked also declined in both periods, but labour quality improved substantially.

Latin America's growth decelerated slightly after 1995, falling from 2.95 to 2.52 per cent. The contribution of labour input was 1.92 per cent before 1995 and 1.89 per cent afterward, accounting for the lion's share of regional growth in both periods. The contribution of capital input rose after 1995 from 0.72 per cent to 0.99 per cent, but remained relatively weak. Mexico's IT investment declined slightly after 1995, while Non-IT investment increased. Nonetheless the contribution of IT investment in Latin America more than doubled, jumping from 0.15 per cent before 1995 to 0.34 per cent afterward or from 21 per cent of the contribution of capital input to 34 per cent. Productivity was essentially flat from 1989 to 2001, rising by 0.31 per cent before 1995 and falling by 0.36 per cent after 1995.

Productivity in sub-Saharan Africa collapsed during 1989–95 but recovered slightly, running at minus 1.63 per cent before 1995 and 0.36 per cent afterward. The contribution of labour input predominated in both periods, but fell from 2.77 per cent to 1.89 per cent, while the contribution of capital input rose from 0.52 per cent to 0.99 per cent. Productivity in North Africa and the Middle East, like that in Latin America, was essentially stationary from 1989 to 2001, falling from a positive rate of 0.50 per cent before 1995 to a negative rate of minus 0.46 per cent afterward.

World Output, Input and Productivity

The final step in our analysis of the world growth resurgence is to describe and characterize the levels of output, input, and productivity for the world economy, the seven economic regions, and the 14 major economies in Table 3. We present levels of output per capita for 1989, before the transition from socialism, 1995, the start of the worldwide IT investment boom, and 2003, the end of the period covered by our study. We also present input per capita and productivity for the years 1989, 1995 and 2003, where productivity is defined as the ratio of output to input.

The G7 economies led the seven economic regions in output per capita, input per capita, and productivity throughout the period 1989–2003. Output per capita in the G7 was, nonetheless, well below US levels. If we take US output per capita in 2000 as 100.0, G7 output per capita was 66.9 in 1989, 72.8 in 1995 and 85.5 in 2003. For comparison: US output per capita was 80.6, 86.3, and 106.4 in these years.

The output gap between the United States and the other G7 economies has widened considerably, especially after 1995. Canada was very close to the United States in output per capita in 1989, but dropped substantially behind by 1995. The United States–Canada gap widened further

Information Technology and the World Economy, Table 3 Levels of output and input per capita and productivity (US = 100 in 2000). The levels for groups and the world are averages weighted by population share

	Output per capita			Input per capita			Productivity		
Region/country	1989	1995	2003	1989	1995	2003	1989	1995	2003
World	**18.9**	**20.0**	**23.9**	**38.5**	**38.5**	**42.4**	**49.0**	**52.0**	**56.3**
G7	66.9	72.8	85.5	72.8	77.4	86.4	91.9	94.1	99.0
Developing Asia	6.0	8.5	12.1	19.1	21.5	26.2	31.7	39.7	46.1
Non-G7	51.5	56.0	68.0	61.9	64.9	75.9	83.2	86.4	89.5
Latin America	18.6	20.0	21.0	27.1	28.2	30.5	68.4	71.0	68.7
Eastern Europe	34.3	22.5	29.3	43.2	41.4	42.6	79.4	54.4	68.8
Sub-Saharan Africa	5.3	4.8	5.0	15.7	15.7	16.7	33.5	30.6	30.0
N. Africa and Middle East	12.5	14.2	17.0	22.3	23.2	27.3	55.9	61.1	62.3
Seven world major economies (G7)									
Canada	79.4	80.2	91.0	75.0	75.7	83.2	105.9	105.9	109.5
France	54.5	57.4	64.7	53.7	57.4	62.1	101.5	100.0	104.2
Germany	59.0	65.5	69.4	71.6	74.3	78.0	82.4	88.2	89.0
Italy	57.7	62.5	69.9	55.9	59.2	70.7	103.2	105.6	98.9
Japan	56.3	64.4	70.8	72.5	78.3	81.7	77.7	82.2	86.7
United Kingdom	56.9	61.8	73.7	61.7	67.5	73.9	92.2	91.6	99.8
United States	80.6	86.3	106.4	84.4	89.1	101.4	95.5	96.9	104.9
All G7	**66.9**	**72.8**	**85.5**	**72.8**	**77.4**	**86.4**	**91.9**	**94.1**	**99.0**
Seven major developing and transition economies (GD7)									
Brazil	19.9	20.5	21.5	29.3	29.8	30.8	67.9	68.7	69.8
China	4.8	8.1	13.4	17.9	20.7	28.0	26.9	39.3	48.0
India	5.0	6.0	8.6	15.9	17.0	19.9	31.2	35.3	43.1
Indonesia	8.3	11.3	12.2	23.7	26.8	29.9	35.3	42.3	40.7
Mexico	22.2	22.3	26.6	28.0	29.7	34.9	79.3	75.3	76.1
Russian Federation	41.8	25.1	33.5	50.0	48.0	47.4	83.6	52.4	70.6
South Korea	24.3	35.8	46.5	37.1	45.4	55.0	65.4	78.9	84.5
All GD7	**9.0**	**10.2**	**14.0**	**24.4**	**24.0**	**28.3**	**36.8**	**42.4**	**49.6**

during the last half of the 1990s. Germany, Japan, Italy, and the UK had similar levels of output per capita throughout 1989–2003, but remained considerably behind North America. France lagged the rest of the G7 in output per capita in 1989 and failed to make up lost ground.

The United States was the leader among the G7 economies in input per capita throughout the period 1989–2003. If we take the United States as 100.0 in 2000, G7 input per capita was 72.8 in 1989, 77.4 in 1995, and 86.4 in 2003, while US input per capita was 84.4, 89.1, and 101.4, respectively. Canada, Germany and Japan were closest to US levels of input per capita with Canada ranking second in 1989 and 2003 and Japan ranking second in 1995. France lagged behind the rest of the G7 in input per capita throughout the period with Italy and the UK only modestly higher.

Productivity in the G7 has remained close to US levels, rising from 91.7 in 1989 to 93.9 in 1995 and 96.7 in 2001, with the United States equal to 100.0 in 2000. Canada was the productivity leader throughout 1989–2003 with Italy and France close behind. The United States occupied fourth place in 1989 and 1995, but rose to second in 2003. Japan made substantial gains in productivity, but lagged behind the other members of the G7 in productivity, while Germany surpassed only Japan.

Differences among the G7 economies in output per capita can be largely explained by differences in input per capita rather than gaps in productivity. The range in output was from 64.7 for France to 106.4 for the United States, while the range in input was from 62.1 for France to 101.4 for the United States. Productivity varied more narrowly

from 86.7 for Japan to 109.5 for Canada with French productivity of 104.2 closely comparable to the United States.

In the economies of Developing Asia output per capita rose spectacularly from 6.0 in 1989 to 8.5 1995 and 12.1 in 2003 with the United States equal to 100.0 in 2000. Levels of output per capita in Asia's largest economies, China and India, remained at 13.4 and 8.6, respectively, in 2003. These vast shortfalls in output per capita relative to the industrialized economies are due mainly to differences in input per capita rather than variations in productivity. Developing Asia's levels of input per capita were 19.1 in 1989, 21.5 in 1995, and 26.2 in 2003, while Asian productivity levels were 31.7, 39.7, and 46.1, respectively.

China made extraordinary gains in output per capita, growing from 4.8 in 1989 to 8.1 in 1995 and 13.4 in 2003 with the United States equal to 100.0 in 2000. India had essentially the same output per capita as China in 1989, but grew less impressively to levels of only 6.0 in 1995 and 8.6 in 2003. China's input per capita – 17.9 in 1989, 20.7 in 1995, and 28.0 in 2001 – exceeded India's throughout the period. India's 31.2 productivity level in 1989 considerably surpassed China's 26.9. China's productivity swelled to 39.3 in 1995, outstripping India's 35.3. China expanded its lead with a productivity level of 48.0 in 2003 by comparison with India's 43.1.

Indonesia and South Korea grew impressively from 1989 to 1995, but fell victim to the Asian financial crisis during the period 1995–2003. Indonesia maintained its lead over India in output per capita, but dropped behind China in 2003. Indonesia led both China and India in input per capita during 1989–2003. Indonesia's productivity level led both China and India in 1995, but fell behind both economies by 2003. South Korea made substantial gains in productivity, achieving a level close to Japan in 2003, while falling considerably short of Japan's impressive input per capita.

The 15 Non-G7 industrialized economies, taken together, had levels of output per capita comparable to Germany, Italy, Japan, and the UK during 1989–2003. Input per capita for the 15 Non-G7 economies was also very close to these four G7 economies. However, productivity for the group was comparable to that of Germany, the second lowest in the G7.

Before the beginning of the transition from socialism in 1989, output per capita in Eastern Europe and the former Soviet Union was 34.3, well above the world economy level of 18.9, with the United States equal to 100.0 in 2000. The economic collapse that accompanied the transition reduced output per capita to 22.5 by 1995, only modestly higher than the world economy level of 20.0. A mild recovery between 1995 and 2003 brought the region back to 29.3, below the level of 1989, but well above the world economy average of 23.9. Input in the region was stagnant at 43.2 in 1989, 41.4 in 1995, and 42.6 in 2003. Productivity collapsed along with output per capita, declining from 79.4 in 1989 to 54.4 in 1995, before climbing back to 68.8 in 2003.

The downturn in output per capita and productivity was especially severe in the economies of the former Soviet Union. Russia's level of output per capita fell from 41.8 in 1989 to 25.1 in 1995 before recovering feebly to 33.5 in 2003. Russian input per capita remained essentially unchanged throughout the period 1989–2003, while productivity mirrored the decline and subsequent recovery in output, falling from a West European level of 83.6 in 1989 to 52.4 in 1995 before recovering to 70.6 in 2003. We conclude that the transition from socialism failed to restore Eastern Europe and the former Soviet Union to pre-transition levels of output and input per capita by 2003, while productivity remained weaker than before the transition.

For the Latin American region output per capita rose from 18.6 to 21.0 during 1989–2003, input per capita rose from 27.1 to 33.0, but productivity was essentially unchanged at about two-thirds of the US level in 2000. The stall in productivity from 1989 to 2003 was pervasive, contrasting sharply with the rise in productivity in the G7 economies, the Non-G7 industrialized economies, and Developing Asia. Nonetheless, Latin America's lagging output per capita was due chiefly to insufficient input per capita, rather than a shortfall in productivity.

Brazil's economic performance has been anaemic at best and has acted as a drag on the growth of Latin America and the world economy. Despite productivity levels comparable to the rest of Latin America, Brazil was unable to generate substantial growth in input per capita. Although Mexico lost ground in productivity between 1989 and 2003, rising input per capita produced gains in output per capita after 1995, despite a slight decline in the contribution of IT investment.

Output and input per capita in sub-Saharan Africa was the lowest in the world throughout the period 1989–2003, but the level of productivity was slightly higher than Developing Asia in 1989. All the economies of North Africa and the Middle East fell short of world average levels of output and input per capita. Output per capita grew slowly but steadily for the region as a whole during 1989–2003, powered by impressive gains in input per capita, but with stagnant productivity.

Methods and Data Sources

To measure capital and labour inputs and the sources of economic growth, we employ the production possibility frontier model of production and the index number methodology for input measurement presented by Jorgenson (2001). For the G7 economies we have updated and revised the data constructed by Jorgenson (2003). For the remaining 103 economies, we rely on two primary sources of data: the Penn World Table (2002) and *World Bank Development Indicators Online* (2004) provide national accounting data for 1950–2003 for all economies in the world except Taiwan. WITSA's *Digital Planet Report* (2002; 2004) gives data on expenditures on IT equipment and software for 70 economies, including the G7. (Other important sources of data include the International Telecommunication Union, ITU, telecommunications indicators, and the UNDP *Human Development* reports.)

US data on investment in IT equipment and software, provided by the Bureau of Economic Analysis (BEA), are the most comprehensive. (The BEA data are described by Grimm, Moulton and Wasshausen, 2004). We use these data as a benchmark in estimating IT investment data for other economies. For the economies included in the *Digital Planet Report* we estimate IT investment from IT expenditures. The *Digital Planet Report* provides expenditure data for computer hardware, software, and telecommunication equipment on an annual basis, beginning in 1992.

Expenditure data from the *Digital Planet Report* are given in current US dollars. However, data are not provided separately for investment and intermediate input and for business, household, and government sectors. We find that the ratio of BEA investment to WITSA expenditure data for the United States is fairly constant for the periods 1981–90 and 1991–2001 for each type of IT equipment and software. Further, data on the global market for telecommunication equipment for 1991–2001, reported by the ITU, confirms that the ratio of investment to total expenditure for the United States is representative of the global market.

We take the ratios of IT investment to IT expenditure for the United States as an estimate of the share of investment to expenditure from the *Digital Planet Report*. We use the penetration rate of IT in each economy to extrapolate the investment levels. This extrapolation is based on the assumption that the increase in real IT investment is proportional to the increase in IT penetration.

Investment in each type of IT equipment and software is calculated as follows:

$$I_{c,A,t} = \eta_{c,A,t} * E_{c,A,t}$$

where $I_{c,A,t}$, $\eta_{c,A,t}$, and $E_{c,A,t}$ are investment, the estimated investment-to-expenditure ratio, and the *Digital Planet Report* expenditures, respectively, for asset A in year t for country c.

The IT expenditures for years prior to 1992 are projected by means of the following model:

$$\ln(Ec_{i\ t-1}) = \beta_0 + \beta_1 \ln(Ec_{it}) + \beta_2 \ln(y_{it-1})$$

where Ec_{it} represents expenditure on IT asset c and the subscripts i and t indicate country i in year t, and y_{it} is GDP per capita. The model

specifies that, for a country i, spending on IT asset c in year $t-1$ can be projected from GDP per capita in that year and spending on asset c in period t.

Given the estimated IT investment flows, we use the perpetual inventory method to estimate IT capital stock. We assume that the geometric depreciation rate is 31.5 per cent and the service life is seven years for computer hardware, 31.5 per cent and five years for software, and 11 per cent and 11 years for telecommunication equipment. Investment in current US dollars for each asset is deflated by the US price index to obtain investment in constant US dollars.

To estimate IT investment for the 66 economies not covered by the *Digital Planet Reports*, we extrapolate the levels of IT capital stock per capita we have estimated for the 70 economies included in these *Reports.* We assume that IT capital stock per capita for the 40 additional economies is proportional to the level of IT penetration. The details are as follows:

For computers we divide the 70 economies included in the *Digital Planet Reports* into ten equal groups, based on the level of personal computer (PC) penetration in 2003. We estimate the current value s^i_{HW} of computer stock per capita in 2003 for an economy i as:

$$s^i_{HW} = s^{-I}_{HW} * \left(P^i_{HW}/\overline{P}^I_{HW}\right),$$

where $\overline{s}^I_{HW}$ is the average value of computer capital per capita in 2001 of Group I for countries included in the *Digital Planet Report,* P^i_{HW} and $\overline{P}^I_{HW}$ are the PC penetration rates of economy i and the average PC penetration of Group I, respectively.

For the economies with data on PC penetration for 1995, we use the growth rates of PC penetration over 1989–2003 to project the current value of computer capital stock per capita backwards. We estimate computer capital stock for each year by multiplying capital stock per capita by population. For economies lacking the data of PC penetration in 1995 and 1989, we estimate computer capital stock by assuming that the growth rates in the two periods, 1995–2003 and 1989–95, are the same as those for the group to which it belongs.

For software capital stock, we divide the 110 countries into ten categories by level of PC penetration in 2003. We subdivide each of these categories into three categories by degree of software piracy, generating 30 groups. (The information on software piracy is based on study conducted by the Business Software Alliance 2003.) We assume that the software capital stock-to-hardware capital stock ratio is constant in each year for each of the 30 groups:

$$s^i_{HW} = \overline{s}^I_{HW} * \left(s^i_{HW}/\overline{s}^I_{HW}\right)$$

where $\overline{s}^I_{HW}$ is the average software capital stock per capita of Subgroup I in 2003. Since the value of computer stock per capita has been estimated for 1995 and 1989, this enables us to estimate the software capital stock per capita for these two years.

Finally, we define the penetration rate for telecommunications equipment as the sum of mainline and mobile telephone penetration rates. These data are available for all 110 economies in all three years – 1989, 1995 and 2003. We have divided these into ten groups by the level of telecommunications equipment penetration for each year. The current value of telecommunications capital stock per capita is estimated as:

$$s^{it}_{TLC} = \overline{s}^{It}_{TLC} * \left(P^{it}_{TLC}/\overline{P}^{It}_{TLC}\right)$$

where $\overline{s}^{It}_{TLC}$ is the average current of telecommunications equipment capital stock per capita in year t of Group I for economies included in the *Digital Planet Reports* and P^{it}_{TLC} and $\overline{P}^{It}_{TLC}$ are the telecommunications equipment penetration rates of economy i and the average penetration rate of Group I in year t.

We employ Gross Fixed Capital Formation for each of the 103 economies provided by the Penn World Table, measured in current US dollars, as the flow of investment. We use the Penn World Table investment deflators to convert these flows into constant US dollars. The constant dollar value of capital stock is estimated by the perpetual inventory method for each of the 103 economies for 1989 and the following years. We assume a depreciation rate of seven per cent and a service life of 30 years.

The current value of the gross capital stock at a year is the product of its constant dollar value and the investment deflator for that year. We estimate the current value of Non-ICT capital stock of an economy for each year by subtracting the current value of IT stock from the current value of capital stock in that year. Given the estimates of the capital stock for each type of asset, we calculate capital input for this stock, using the methodology presented by Jorgenson (2001).

Finally, labour input is the product of hours worked and labour quality:

$$L_t = H_t * q_t$$

where L_t, H_t, and q_t, respectively, are the labour input, the hours worked, and labour quality. A labour quality index requires data on education and hours worked for each of category of workers.

We extrapolate the labour quality indexes for the G7 economies by means of the following model:

$$\begin{aligned} q_{i,t} = \quad & \beta_0 + \beta_1 \ \text{Education}_{i,t} \\ & + \beta_2 \ \text{Institution} \ 1_i + \beta_3 \ \text{Institution} \ 2_i \\ & + \beta_4 \ \text{Income} \ 1989_i + \beta_5 T \end{aligned}$$

where subscripts i and t indicate economy i in year t. Education is the educational attainment of the population aged 25 or over from the data-set constructed by Barro and Lee (2001). Institution 1 = 'Rule of Law' and Institution 2 = 'Regulatory Quality' are constructed by Kaufmann et al. (2004) for the World Bank; Income 1990 is GDP per capita for 1990 from the Penn World Table and T is a time dummy.

Labour quality is largely explained by educational attainment, institutional quality and living conditions. The model fits well ($R^2 = 0.973$) and all the explanatory variables are statistically significant. We assume that hours worked per worker is constant at 2000 hours per year, so that growth rates of hours worked are the same as employment.

Summary and Conclusions

World economic growth, led by the industrialized economies and Developing Asia, experienced a strong resurgence after 1995. Developing Asia accounted for an astonishing 60 per cent of world economic growth before 1995 and 40 per cent afterward, with China alone responsible for half of this, but output per capita remained well below the world average. Sub-Saharan Africa and North Africa and the Middle East languished far below the world average. Eastern Europe and the former Soviet Union lost enormous ground during the transition from socialism and have yet to recover completely.

The growth trends most apparent in the United States have counterparts throughout the world. Investment in tangible assets, including IT equipment and software, was the most important source of growth. However, Non-IT investment predominated. The contribution of labour input was next in magnitude with labour quality dominant before 1995 and hours worked afterward. Finally, productivity was the least important of the three sources of growth, except during the Asian Miracle before 1995.

The leading role of IT investment in the acceleration of growth in the G7 economies is especially pronounced in the United States, where IT is coming to dominate the contribution of capital input. The contribution of labour input predominated in the Non-G7 industrialized economies, as well as Latin America, Eastern Europe, sub-Saharan Africa, and North Africa and the Middle East. Productivity growth was the important source of growth in Developing Asia before 1995, but assumed a subordinate role after 1995. Productivity has been stagnant or declining in Latin America, Eastern Europe, sub-Saharan Africa, and North Africa and the Middle East.

All seven regions of the world economy experienced a surge in investment in IT equipment and software after 1995. The impact of IT investment on economic growth has been most striking in the G7 economies. The rush in IT investment was especially conspicuous in the United States, but jumps in the contribution of IT capital input in Canada, Japan, and the UK were only slightly lower. France, Germany and Italy also experienced a surge in IT investment, but lagged considerably behind the leaders. While IT investment

followed similar patterns in the G7 economies, Non-IT investment varied considerably and explains important differences among growth rates.

Although the surge in investment in IT equipment and software is a global phenomenon, the variation in the contribution of this investment has increased considerably since 1995. Following the G7, the next most important increase was in Developing Asia, led by China. The Non-G7 industrialized economies followed Developing Asia. The role of IT investment more than doubled after 1995 in Latin America, Eastern Europe, and North Africa and the Middle East, and sub-Saharan Africa.

See Also

- ▶ Production Functions
- ▶ Technical Change
- ▶ Technology

Acknowledgments The Economic and Social Research Institute provided financial support for work on the G7 economies from its program on international collaboration through the Nomura Research Institute. Alessandra Colecchia, Mun S. Ho, Kazuyuki Motohashi, Koji Nomura, Jon Samuels, Kevin J. Stiroh, Marcel Timmer, and Bart van Ark provided valuable data. The Bureau of Economic Analysis and the Bureau of Labor Statistics assisted with data for the United States and Statistics Canada contributed the data for Canada. We are grateful to all of them but retain sole responsibility for any remaining deficiencies.

Bibliography

Ahmad, N., P. Schreyer, and A. Wolfl. 2004. ICT investment in OECD countries and its economic impact. In *The economic impact of ICT: Measurement, evidence, and implications*, ed. OECD. Paris: OECD.

Baldwin, J.R., and T.M. Harchaoui. 2003. *Productivity growth in Canada – 2002*. Ottawa: Statistics Canada.

Barro, R.J., and J.-W. Lee. 2001. International data on educational attainment: Updates and implications. *Oxford Economic Papers* 53: 541–563.

Business Software Alliance. 2003. *Global software piracy study: Trends in piracy, 1994–2002.* Washington, DC: Business Software Alliance USA, June. Online. Available at http://www.caast.com/resources/2003_global_study.pdf. Accessed 22 Apr 2007.

Daveri, F. 2002. The new economy in Europe: 1992–2001. *Oxford Review of Economic Policy* 18: 345–362.

Grimm, B., B. Moulton, and D. Wasshausen. 2005. Information processing equipment and software in the national accounts. In *Measuring capital in the new economy*, ed. C. Corrado, J. Haltiwanger, and D. Sichel. Chicago: University of Chicago Press.

Heston, A., R. Summers, and B. Aten. 2002. *Penn world table version 6.1*. Philadelphia: Center for International Comparisons at the University of Pennsylvania (CICUP), October. Online. Available at http://pwt.econ.upenn.edu/aboutpwt.html. Accessed 1 Feb 2007.

International Telecommunications Union. 2004. *World telecommunications indicators.* Geneva: International Telecommunications Union, October. Online. Available at http://www.itu.int/ITUD/ict/publications/world/world.html. Accessed 1 Feb 2007.

Jorgenson, D.W. 2001. Information technology and the U.S. economy. *American Economic Review* 91: 1–32.

Jorgenson, D.W. 2003. Information technology and the G7 economies. *World Economics* 4(4): 139–170.

Jorgenson, D.W. 2005. Accounting for growth in the information age. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Jorgenson, D.W., and K. Motohashi. 2005. Information technology and the Japanese economy. *Journal of the Japanese and International Economies* 19: 460–481.

Jorgenson, D.W., and K.J. Stiroh. 2000. Raising the speed limit: U.S. economic growth in the information age. *Brookings Papers on Economic Activity* 2000(1): 125–211.

Jorgenson, D.W., and E. Yip. 2000. Whatever happened to productivity growth? In *New developments in productivity analysis*, ed. C.R. Hulten, E.R. Dean, and M.J. Harper. Chicago: University of Chicago Press.

Jorgenson, D.W., M.S. Ho, and K.J. Stiroh. 2005. *Information technology and the American growth resurgence*. Cambridge, MA: MIT Press.

Kaufmann, D., A. Kray, and M. Mastruzzi. 2004. *Governance matters III: Governance indicators for 1996–2002*. Washington, DC: World Bank.

Krugman, P. 1994. The myth of Asia's miracle. *Foreign Affairs* 73(6): 62–78.

Lau, L.J. 1999. The sources of East Asian economic growth. In *The political economy of comparative development in the 21st century*, ed. G. Ranis, S.-C. Hu, and Y.-P. Chu. Northampton: Edward Elgar.

Maddison, A. 2001. *The world economy: A millennial perspective*. Paris: OECD.

OECD (Organisation for Economic Co-operation and Development). 2002. *Purchasing power parities and real expenditures, 1999 benchmark year*. Paris: OECD.

Oliner, S.D., and D.J. Sichel. 2000. The resurgence of growth in the late 1990s: Is information technology the story? *Journal of Economic Perspectives* 14(4): 3–22.

Piatkowski, M., and B. van Ark. 2005. ICT and productivity growth in transition economies: Two-phase

I

convergence and structural reforms. TIGER Working Paper Series No. 72, Warsaw, January.
UNDP. 2004. *Human development report 2004*. New York: United Nations.
van Ark, B., J. Melka, N. Mulder, M. Timmer, and G. Ypma. 2003, updated 2005. ICT investment and growth accounts for the European Union, *1980–2000.* Brussels: European Commission.
WITSA (World Information Technology and Services Alliance). 2002. *Digital planet report*. Washington, DC: WITSA.
World Bank. 2004. *World development indicators 2004*. Washington, DC: World Bank.
Young, A. 1995. The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience. *Quarterly Journal of Economics* 106: 641–680.
Young, A. 2003. Gold into base metals: Productivity growth in the People's Republic of China during the reform period. *Journal of Political Economy* 111: 1220–1261.

Information Theory

Esfandiar Maasoumi

Information theory is a branch of mathematical statistics and probability theory. Thus, it can and has been applied in many fields, including economics, that rely on statistical analysis. As we are concerned with it, the technical concept of 'information' must be distinguished from the semantic concept in common parlance. The simplest and still the most widely used technical definitions of information were first introduced (independently) by Shannon and Wiener in 1948 in connection with communication theory. Though decisively and directly related, these definitions must also be distinguished from the definition of 'information' introduced by R.A. Fisher in 1925 for estimation theory.

Whenever observations are made or experiments conducted, we seek information about the underlying populations. Information theory provides concepts and definitions by means of which we may measure formally what can be inferred from the sampled data. More narrowly interpreted, we may view these concepts and definitions as summary statistics representing the associated (empirical) distributions, much like the moments of distributions. These concepts, however, also admit the intuitive meanings not unlike the notions of 'information' and 'uncertainty' in common parlance. One such concept, the *entropy* of a distribution, is central to information theory and is a measure of disorder or uncertainty. It was the definition of entropy that first caught the attention of Henri Theil and led to the use of information theory in *economics* in the 1960s. Following this discovery two measures of income inequality, measures of divergence and/or concentration in trade and industry, and many other economic applications were introduced in Theil (1967). Further applications in economic theory and other social sciences are discussed in Theil (1972, 1980).

A somewhat different but equally important set of developments of information theory have taken place in *econometrics*. Information criteria exist which measure the 'divergence' between populations. The use of such criteria helps to discriminate statistically between hypotheses, select models and evaluate their forecasting performance. These are essential steps in model evaluation and inference in econometrics.

Given a sample and (possibly) some prior information, a so-called 'maximum entropy (ME)' distribution of the underlying continuous process may be derived. This distribution and its quantiles and moments can be used in order to resolve a number of important problems including that of undersized samples. Theil and Laitinen (1980) formulated the ME distribution, and Theil and Fiebig (1984), give a complete coverage of this topic.

More general concepts of entropy and the associated measures of divergence can be used to develop a family of Generalized Entropy (GE) measures of inequality and in the area of 'multidimensional' welfare analysis. A by-product of this is the development of a 'summary welfare attribute' which serves as an index of several miscellaneous attributes such as income, wealth and physical quality of life indices. Another by-product is the development of 'information

efficient' functional forms when the regression function is unknown *a priori*.

Applications to the classical theory of consumer and producer demand equations is noteworthy. The feature of this theory that provides the connection with information theory is the concept of 'value shares', the expenditure on a given commodity (input factor) divided by total expenditure. Thus the value shares have the same mathematical properties as probabilities ($0 \leq p_i \leq 1$). Theil (1967) noted some of the first attempts by Lisman (1949) and Pikler (1951) to draw analogies between econometric and information theory concepts. Emphasizing a point elaborated upon by De Jongh (1952), Theil wrote: 'the reason information theory is nevertheless important in economics is that it is more than a theory dealing with information concepts. It is actually a general partitioning theory in the sense that it presents measures for the way in which some set is divided into subsets ….' And then, 'it may amount to dividing certainty (probability 1) into various possibilities none of which is certain, but it may also be an allocation problem in economics.'

Our account begins with the introduction of some basic concepts.

Central Definitions and Concepts

An extensive literature exists which treats information theory in an axiomatic manner, much of it stimulated by the work of Shannon (1948) and Wiener's suggestive remarks in Wiener (1948). Kullback (1959) provides a comprehensive bibliography. From Stumpers (1953) it is clear, however, that some important contributions had appeared prior to 1948. For our purposes it will suffice to note that the occurrence of an event contains (or conveys) information. Thus one needs 'information functions' that measure the amount of information conveyed. If, as is usual, we are concerned with random events, there usually exist prior probabilities of occurrence of events and posterior probabilities. Hence, given an information function, we may measure the 'information gain' as between the prior and the posterior probabilities. Further, given an experiment and a *set* of observations, we may measure the 'expected information gain'.

The most widely used information function is $-\log p_i$, where p_i denotes the probability of a random variable X taking a value x_i. This function is non-negative and satisfies the 'additivity axiom' for *independent events*. That is, the information that both of two independent events have occurred is equal to the sum of the information in the occurrence of each of the events. More general information functions will be discussed later. The base of the logarithm (usually e or 2) determines the desired 'information unit', 'bits' when 2 is the base, and 'nits' when the natural base is employed. It may be useful to list some of the axioms that are employed to restrict the form of information functions:

Let such functions be denoted by $h(\cdot)$.

Axiom 1 – $h(\cdot)$ is a function only of the probability of events (p, say).
Axiom 2 – $h(p)$ is continuous in p, $0 < p \leqslant 1$.
Axiom 3 – $h(p)$ is a monotonically decreasing function of p.

This last axiom is quite intuitive. For instance, when we receive a definite message (observation) that a most unlikely event (p close to zero) has occurred, we are more highly surprised (informed) than if the event had a high probability (p close to 1) of occurrence. Thus, we may impose the following restrictions:

$$h(0) = \infty, \quad h(1) = 0 \quad \text{and} \quad h(p_1) > h(p_2), \quad \text{if } 0 \leq p_1 < p_2 \leq 1.$$

There are many functions that satisfy these axioms. $\log 1/p = -\log p$ is uniquely identified (apart from its base) on the basis of a further axiom, the *additivity* of $h(.)$ in the case of independent events.

A generalization of the above definition is needed for general situations in which the messages received are not completely reliable. An important example is the case of forecasts of weather or economic conditions. Suppose p_1 is

the probability of occurrence of an event (perhaps calculated on the basis of frequency of occurrence in the past), and p_2 the probability of the same event *given that it had been predicted to occur* (calculated in the same manner). We may then enquire as to the merit or the 'information gain' of such predictions. Given the observation that the predicted event has occurred, this information gain is defined as follows:

$$h(p_1) - h(p_2) = \log p_2/p_1$$

More generally, the information content of data which may be used to 'update' prior probabilities is defined by:

Information gain (IG) = log (posterior probability/prior probability). Since typically, data are subject to sampling variability, the 'expected information gain' is the appropriate measure of information gain. Thus, the expected value of IG defined above may be obtained with respect to either the prior or the posterior distributions over the range of the possible values for the random event. This gain is the difference between the 'expected information' in the two distributions. Consider a set of n mutually exclusive and exhaustive events $E_1, \ldots, E_n$ with corresponding probabilities $p_1, \ldots, p_n$ ($\Sigma p_n = 1$, $p_i \geq 0$). Then occurrence of an event contains $h(p_i)$ information with probability p_i. *Before* any observation is made, the 'expected information' in the distribution $p = (p_1, \ldots, p_n)$ is given by:

$$0 \leq \sum_{i=1}^{n} p_i h(p_i) = H(p) \leq \log n. \qquad (1)$$

When $h(p_i) = -\log p_i$, the convention: $p_i \log p_i = 0$ if $p_i = 0$, is used. The maximum, $\log n$, occurs when $p_i = 1/n$ for all $i = 1, \ldots, n$. This is a situation of maximum prior 'uncertainty' or 'disorder' in the distribution p. Contrast this with $p_1 = 1, p_i = 0 \forall i \neq 1$, in which case $H(p) = 0$.

A dual concept to 'expected information' may be defined. This is the 'Entropy' which measures the 'uncertainty' or 'disorder' in a distribution p as defined by $H(p)$ in equation (1). Entropy is a central concept in information theory and its applications. It is also considered as an index of how close a distribution is to the uniform distribution. Note that an otherwise unknown distribution (p) may be determined by maximizing its entropy subject to any available restrictions. For instance the first few moments of the distribution may be specified *a priori*. The distribution so obtained is called the 'Maximum Entropy' (ME) distribution. In Theil and Laitinen (1980), *continuity* as well as the first moment of the observed data are utilized in order to obtain the ME distribution of the data and its higher moments.

The concept of 'divergence' or 'distance' between distributions naturally follows those of expected information gain and entropy. Instead of the prior and posterior distributions referred to earlier, it may be more suggestive to consider competing distributions (perhaps resulting from competing hypotheses), $f(x)$ and $g(x)$, which may generate a random variable x with the range denoted by R. There are two *directional* measures of divergence between $f(\cdot)$ and $g(\cdot)$. These are:

$$I(2,1) = \int_R g(x) \log \frac{g(x)}{f(x)} \mathrm{d}x$$

and

$$I(1,2) = \int_R f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x.$$

A *non-directional* measure in the same context may be an average of the two directional criteria given above. For instance:

$$J(1,2) = \int_R [\log f(x) - \log g(x)][f(x) - g(x)]\mathrm{d}x$$
$$= (1,2) + I(2,1).$$

This is the well known Kullback–Leibler information criterion used extensively in many applications.

Various generalizations of these criteria are obtained either by generalizations of the form of the information function (which typically include the logarithmic form as a special case), or by generalizations of the *metrics* that include $J(1,2)$ above.

Some properties of these central concepts and their generalizations will be discussed in the following sections. It will be illuminating, however, to close this section by demonstrating an interesting connection between the information concepts defined so far and an important definition given by R.A. Fisher (1925):

Let x be a random variable taking values in the space S with the p.d.f. $f(., \boldsymbol{\Theta})$ with respect to a σ-finite measure υ. Assume $f(\cdot, \boldsymbol{\Theta})$ differentiable w.r.t. $\boldsymbol{\Theta}$ and:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Theta}} \int_c f(x, \boldsymbol{\Theta}) \ \mathrm{d}\upsilon = \int_c \frac{\mathrm{d}\ f(x, \boldsymbol{\Theta})}{\mathrm{d}\boldsymbol{\Theta}} \ \mathrm{d}\upsilon$$

for any measurable set $c \in S$. Fisher defined the following measure of information on $\boldsymbol{\Theta}$ contained in x:

$$\phi(\boldsymbol{\Theta}) = E\left(\frac{\mathrm{d} \log f}{\mathrm{d}\boldsymbol{\Theta}}\right)^2 = V\left(\frac{\mathrm{d} \log f}{\mathrm{d}\boldsymbol{\Theta}}\right)$$

where V denotes the variance when $E(\mathrm{d} \log f/\mathrm{d}\boldsymbol{\Theta}) = 0$.

If there is a unique observation (of x) with probability 1 corresponding to *each value of* $\boldsymbol{\Theta}$, then the random variable (i.e. its distribution) has the maximum information. The least information exists if the random variable has the *same distribution* for all $\boldsymbol{\Theta}$. Thus, one might measure the sensitiveness of x with respect to $\boldsymbol{\Theta}$ by the extent to which its distribution changes in response to (infinitesimal) changes in $\boldsymbol{\Theta}$. If $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}' = \boldsymbol{\Theta} + \delta\boldsymbol{\Theta}$ are two values of $\boldsymbol{\Theta}$, a suitable measure of 'distance' or 'divergence', $D[f(\boldsymbol{\Theta}), f(\boldsymbol{\Theta}')]$, is required. An example of $D[\cdot]$ was given earlier, and many more criteria have been proposed. It may be shown that many such criteria are *increasing* functions of Fisher's information $[\phi(\boldsymbol{\Theta}) \geq 0]$. To give an example, consider the Hellinger distance:

$$\cos^{-1} \int \left[f(x, \Theta) \cdot f(x, \Theta')\right]^{1/2} \ \mathrm{d}\upsilon.$$

Using a Taylor expansion of $f(x, \boldsymbol{\Theta}')$ and neglecting terms of power 3 or more in $\delta\boldsymbol{\Theta}$, we find:

$$\cos^{-1} \int f(\boldsymbol{\Theta}) \left\{1 - \frac{1}{8}\left[\frac{f'(\boldsymbol{\Theta})}{f(\boldsymbol{\Theta})}\right]^2 (\delta\boldsymbol{\Theta})^2\right\} \mathrm{d}\upsilon$$
$$= \cos^{-1}\left[1 - \frac{1}{8}\phi(\boldsymbol{\Theta})(\delta\boldsymbol{\Theta})^2\right]$$

where $\phi(\boldsymbol{\Theta})$ is indeed Fisher's information. Thus, such divergence criteria and $\phi(\boldsymbol{\Theta})$ are equivalent measures of the sensitivity of the random variables (p.d.f., s) with respect to small changes in the parameter values. These observations provide a basis for measuring the distance between competing hypotheses (on $\boldsymbol{\Theta}$) and model selection techniques in econometrics.

Applications in Economics and Econometrics

Measurement of Economic Inequality

Theil (1967) observed that the entropy, $H(y)$, was a remarkably useful measure of 'equality'. If $y = (y_1, \ldots, y_N)$ denotes the non-negative income *shares* of N individuals, the entropy of y, by definition, measures its distance from the rectangular distribution, $y_i = 1/N$, which is the case of complete equality. Thus, the difference between $H(y)$ and its maximum value, $\log N$, may be used as a measure of *inequality*. This measure satisfies the 'three fundamental welfare requirements', namely symmetry (S), Homogeneity (H), and the Pigou–Dalton principle of transfers (PT). In addition, Theil (1967) demonstrated extremely useful *additive decomposability* properties of this measure. In recent axiomatic treatments by Bourguignon (1979), Shorrocks (1980), Cowell and Kuga (1981) and Foster (1983), the following question is posed: what is a suitable measure of inequality among general classes of functional forms which are restricted to satisfy the above three requirements in addition to Theil's decomposability? The last condition identifies Theil's first measure, $\mathrm{T1} = \log N - H(y)$, and a second information measure proposed in Theil (1967). The latter is defined as follows:

$$\mathrm{T2} = -\log N - \frac{1}{N}\sum \log y_i.$$

The choice between these two measures implies preferences that may be formulated by Social Welfare Functions (SWF). From a practical viewpoint, however, the choice between T1 and T2 may also be made on the basis of their decomposability properties. These may be briefly described as follows:

Let there by G exclusive sets (groups) of individuals, $S_1, \ldots, S_G$, with N_g denoting the number of individuals in S_g, $g = 1, \ldots, G$ ($\Sigma N_g = N$). Let y_g be the *share* of S_g in total income. Then we have:

$$\mathrm{T1}(y) = \sum_{g=1}^{G} y_g \log \frac{y_g}{N_g/N} + \sum_{g} y_g \left[\log N_g - H_g(y)\right].$$

Here, $H_g(y)$ denotes the entropy of group g calculated from (y_i/y_g) for all $i \in S_g$. The first term above is the 'between-group' inequality, and the second term is a weighted average of the 'within-group' inequalities (the term in the square brackets). This decomposition is essential in analysing the incidence of inequality amongst the population subgroups (e.g. defined by age, race, region, etc.)

A similar decomposition formula holds for the second measure T2. The major difference is that the 'within-group' inequalities in T2 are weighted by the groups' *population shares* (N_g/N) rather than their income shares (y_g). The decomposition for T2 is somewhat preferable to that for T1 (see Shorrocks 1980) as it permits a less ambiguous discussion of such questions as: what is the contribution of the inequality in the gth group to total inequality? This is partly because y_g are sensitive to redistributions (distributional changes) whereas population shares (N_g/N) are not so by design.

A generalization of the concept of information functions and the entropy has been employed by Toyoda, Cowell and Kuga to define the family of Generalized Entropy (GE) inequality indices. The GE is defined as follows:

$$\mu_\gamma(y) = \frac{1}{N} \sum_i \left[(Ny_i)^{1+\gamma} - 1\right] / \gamma(\gamma + 1)$$

where $\mu_0()$ and $\mu_{-1}()$ are, respectively, the T1 and T2 defined earlier. $\gamma \leq 0$ ensures the convexity of GE members, and for values of $\gamma \ll 0$ the GE is ordinally equivalent to the class of measures proposed earlier by Atkinson (1970) by direct reference to the SWFs. $-\gamma = \upsilon \geq 0$ is referred to as the 'degree of inequality aversion' exhibited by the underlying SWF. The information function underlying the GE indices is $-1/\gamma\left(y_i^{\gamma} - 1\right)$ which includes $-\log y_i$ as a special case. The generalized entropy corresponding to this function is given by:

$$H_\gamma(y) = \sum_i \frac{1}{\gamma(\gamma + 1)} \left[(Ny_i)^{\gamma} - 1\right] \cdot y_i.$$

The GE measures are also decomposable in the manner described above.

Multi-Dimensional Welfare Analysis

The recognition that welfare depends on more than any single attribute (e.g. income) has lead to analyses of welfare functions and inequality in the multi-dimensioned space of several attributes, such as incomes (and its factor components), wealth, quality of life and basic needs indices. Pioneering work that deals directly with individual utilities (as functions of these attributes) and Social Welfare Functions is primarily due to Kolm (1977) and Atkinson and Bourguignon (1982). The measurement approach due to Maasoumi (1986a) poses the same question statistically and, surprisingly, provides a pure measurement (index number) interpretation of the SWF approach with equivalent solutions. The measurement approach seeks a 'summary share' or an index which may be employed to represent the miscellaneous attributes of interest. Information theory is utilized to obtain 'summary share' distributions without explicitly imposing any restrictive structure on preferences and behaviour. Noting that a measure of inequality is a summary statistic for a distribution (much as the moments of a p.d.f.). Maasoumi (1986a) poses the following question: Which summary distribution (index) is the 'closest' to the distribution of the welfare attributes of interest? Given suitable information

criteria for measuring the 'distance' between distributions, one can find a summary (or representative) *share vector* (distribution) that minimizes this distance. Briefly: let y_{if} be the share of the ith individual, $i \in [1, N]$, from the fth attribute, $f \in [1, M]$. There are M distributions, $y_f = (y_{1f}, \ldots, y_{Nf})$, and we seek a distribution, $S = (S_1, \ldots, S_N)$, which is closest to these M distributions as measured by the following generalized information measure of divergence:

$$D(\gamma) = \frac{1}{\gamma(\gamma+1)} \sum_{f=1}^{M} \alpha_f \sum_{i=1}^{N} S_i \left[\left(S_i / y_{if} \right)^{\gamma} - 1 \right]$$

where α_f is the weight given to the fth attribute. Minimizing $D(\gamma)$ with respect to S_i subject to $\Sigma S_i = 1$, we find:

$$S_i \propto \left(\sum_f \delta_f y_{if}^{-\gamma} \right)^{-1/\gamma}, \quad \sum_f \delta_f = 1.$$

This is a functional of the CES variety which includes the Cobb–Douglas ($\gamma = 0$) and the linear ($\gamma = -1$) forms as special cases. The S_i may be regarded as individual utility functions with the optimal distributional characteristics implied from $D(\gamma)$.

Once $S = (S_1, \ldots, S_N)$ is so determined, any of the existing measures of inequality may be used to measure multivariate inequality represented in S. Members of GE have been used for this purpose in Maasoumi (1986a). It may be shown that only two members of GE, Theil's T1 and T2, and some values of γ, provide fully decomposable measures of multi-dimensional inequality. Full decomposition refers to decomposition by population subgroups as well as by the inequality in individual welfare attributes. An interesting alternative method, Principal Components, is seen to be a special case of the above approach. It corresponds to the case $\gamma = -1$, with α_f being the elements of the first characteristic vector of $y'y, y = (y_{if})$ being the distribution matrix. This feature of information theory is not surprising. As S. Kullback and others have noted, one of its great advantages is the generality that it affords in the analysis of statistical issues, with the suggestion of new solutions and useful interpretations of the old.

'Information Efficient' Functional Forms

Economic theory is generally silent on the specific form of functional relationships between variables of interest. Certain restrictions on the general characteristics of such relations are typically available, but ideally one must use the available data to determine the appropriate functional form as well as its parameters. The common practice in econometrics is either to specify flexible functional families, or to test specific functional forms for statistical adequacy (e.g. see Judge et al 1980, ch. 11). The criteria of section "Multi-Dimensional Welfare Analysis" may be used, however, to obtain functions of the data with distributions which most closely resemble the empirical distribution of the observations. For instance, using $D(\gamma)$ from the previous section, let $S_i = f(x_i)$ be the indeterminate functional relationship between the variables $x_i = (x_{1i}, \ldots, x_{ki})$ at each observation point $i = 1, \ldots, T$. The variable set x_i may or may not include the endogenous (dependent) variable in an explicit regression context. Maasoumi (1986b) shows that, in the latter case, the CES functional form is 'ideal' according to $D(\gamma)$, and in the former the usual Box–Cox transformation is obtained. According to $D(\gamma)$, any other functional forms will distort the distributional information in the sample. The value of $D(\gamma)$ for any approximate regression function, less its minimum, is an interesting measure of the informational inefficiency of that regression function.

Tests of Hypotheses and Model Selection

Sample estimates of the measures of divergence $I(1, 2)$, $I(2, 1)$ and $J(1, 2)$ defined above may be used to test hypotheses or to choose the 'best' models. For instance, the minimum value of $I(1, 2)$, denoted by $I(*:2)$ and called 'the minimum discrimination information', is obtained for a given distribution $f_2(x)$ with respect to all $f_1(x)$, such that

$$\int T(x) f_1(x) dx = \boldsymbol{\Theta}$$

where $\boldsymbol{\Theta}$ are constants and $T(x)$ are measureable statistics. For example, $T(x) = x$ and $\boldsymbol{\Theta} = \mu$, restricts the mean of possible distributions $f_1(x)$. General solutions for $f_1(x)$ [denoted $f^*(x)$] and $I(*{:}2)$ are given in Kullback (1959, ch. 3) and elsewhere.

$I(*{:}2)$, and the corresponding values for $I(1{:}*)$ and $J(*)$, may be estimated by replacing $\boldsymbol{\Theta}$ (and the other unknown parameters) by their sample estimates ($\widehat{\boldsymbol{\Theta}}$) when $f_2(x)$ is the generalized density of n independent observations. Given a sample 0_n, we denote this estimate by $\widehat{I}(* : \ 2; 0_n)$. This statistic measures the minimum discrimination information between a population with density $f^*(x)$ (with $\Theta = \widehat{\boldsymbol{\Theta}}$ etc.), and the population with the density $f_2(x)$. The justification for its use in tests of hypotheses and model selection is that, the non-negative statistic $\widehat{I}(\cdot)$ is zero when $\widehat{\boldsymbol{\Theta}}$ is equal to $\boldsymbol{\Theta}$ of the population with density $f_2(x)$, becoming larger the worse is the resemblance between the sample and the hypothesized population $f_2(x)$.

To illustrate, consider the linear regression model

$$y = X\beta + U$$

where $U \sim N\ (0,\ \Sigma)$ and $X \sim T \times K$ and of rank K.

Consider two competing hypotheses:

$$H_1 : \beta = \beta(\text{no restriction}), \qquad H_2 : \beta = \beta^2.$$

Then, it may be verified that:

$$J(1,2) = 2I(1,2) = \left(X\beta - X\beta^2\right)' \Sigma^{-1} \left(X\beta - X\beta^2\right)$$
$$= \frac{1}{\sigma^2} \left(X\beta - X\beta^2\right)' \left(X\beta - X\beta^2\right).$$
$$\text{if } \Sigma = \sigma^2 I = \left(\beta - \beta^2\right)' (X'X) \left(\beta - \beta^2\right) / \sigma^2.$$

Replacing the unknown parameters, $\boldsymbol{\Theta} = (\beta, \sigma^2)$, with their respective unbiased OLS estimates, we find:

$$\widehat{J}(H_1; H_2) = \left(\widehat{\beta} - \beta^2\right)' (X'X) \left(\widehat{\beta} - \beta^2\right) / \widehat{\sigma}^2.$$

And if $\beta^2 = 0$:

$$\widehat{J}(H_1, H_2) = \widehat{\beta}' (X'X) \widehat{\beta} / \widehat{\sigma}^2$$

which may be recognized as proportional to Hotelling's T^2 statistic with an F distribution with K and $T - K$ degrees of freedom.

The above example produced a statistic with a known finite sample distribution. In this situation, we are in effect rejecting H_2 if:

$$\text{Prob}\left\{\widehat{I}(* : H_2) - \widehat{I}(*H_1) \geq C | H_2\right\} \leq \alpha$$

where C is chosen to control the size (α) and the power of the test.

More generally, when the exact distributions are not known, asymptotic procedures may be employed. Suppose, for instance, that the competing populations (hypotheses), H, are members of the exponential family [which includes $f^*(x)$]. Let the admissible range of parameter values be denoted by Ω, and the range (value) specified by $H_i \in H$ denoted by ω_i. It may be shown that (see Kullback 1959):

$$\widehat{I}(* : H) = \log \left[\max_{\Omega} f*(x) / \max_{\omega_i} f*(x) \right]$$
$$= -\log \lambda_i$$

where $_i$ is the Neyman–Pearson (likelihood-ratio) statistic. Under certain regularity conditions, the statistic $-2 \log \lambda_i$ is asymptotically distributed as X^2. Also, in the same situation, for two competing hypotheses H_1: $\boldsymbol{\Theta} \in \omega_1$, H_1: $\boldsymbol{\Theta} \in \omega_2$ it may be shown that:

$$\widehat{I}(* : H_2) - \widehat{I}(* : H_1) = -\log \lambda*$$

where $\lambda^* = \max_{\boldsymbol{\Theta} \in \omega 2} f^*(x) / \max_{\boldsymbol{\Theta} \in \omega 1} f^*(x)$. Variants of such statistics are also useful for tests of 'non-nested' hypotheses. For the distribution of $\lambda*$ and its extensions see (e.g.) Chernoff (1954) and Cox (1961). Finally, we note that the above test reduces to $\widehat{I}(* : H_2)$ when H_2: $\boldsymbol{\Theta} \in \omega$ and H_1: $\boldsymbol{\Theta} \in \Omega - \omega$. In such cases $\widehat{I}(* : H_1) = 0$.

Variations to the information criteria described above have been proposed for 'model selection' in econometrics. Akaike (1973) proposed a measure

based on the Kullback–Leibler criterion. We give this criterion (AIC) for the problem of choosing an optimal set of regressors in the standard regression model. As before, let $y = X\beta + U, U \sim N(0, \sigma_u^2 I)$, be the 'comprehensive' model, with $X = [X_1, X_2]$ and $\beta = (\beta'_1, \beta'_2)'$ representing a full rank partition $K_1 + K_2 = K$. Under the null hypothesis $R\beta = [0, I_{k2}]\beta = 0$ ($\beta_2 = 0$), the AIC is as follows:

$$\begin{aligned} \text{AIC} &= -\frac{2}{T}\log l(y, \beta) + 2K_1/T \\ &= \log(y'M_1 y/T) + 2K_1/T \end{aligned}$$

where $l()$ is the likelihood function and $M_1 = I - X(X'_1X_1)X'_1$. One proceeds to choose K_1 (and hence M_1) so as to minimize AIC. The first term decreases with K_1, thus the above criterion incorporates the trade-off between parsimony and 'fit' of a model. If one proceeds as though σ^2 were known in the likelihood function. Amemiya (1976) shows that:

$$\text{AIC}\ \left(\sigma^2\ \text{known}\right) = y'M_1 y/T + \sigma^2(2K_1)/T$$

An estimate of this last criterion may be based on the unbiased OLS estimates of σ^2 with or without the restrictions ($\beta_2 = 0$). The latter estimate is equivalent to the so-called 'Cp criterion', and the former is equivalent to the 'Prediction Criterion' for model selection. For other variations to AIC (e.g. Sawa's BIC criterion) see Judge et al. (1980, Section 11.5).

Conclusions

The above examples do not do justice to a remarkable range of currently available applications of the information criteria in economics and econometric inferences. We hope, however, that they suffice to show: (1) the usefulness of the general approach in encompassing many different and often ad hoc procedures in econometric inference; and (2) how new methods with plausible and intuitive appeal may be derived in order to resolve many hitherto unresolved problems. The full potential for further applications of information theory and its formal discipline in economic and econometric theory is great. The current level of interest in this potential is extremely promising.

See Also

- ▶ Entropy
- ▶ Hypothesis Testing
- ▶ Prediction
- ▶ Signalling

Bibliography

Akaike, H. 1973. Information theory and the extension of the maximum likelihood principle. In *International symposium on information theory*, ed. B.N. Petrov and F. Csaki. Budapest: Akadémiaikiado.

Amemiya, T. 1976. *Selection of regressors*, Technical Report, vol. 225. Stanford: Stanford University.

Atkinson, A.B.. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Atkinson, A.B.., and F. Bourguignon. 1982. The comparison of multi-dimensional distributions of economic status. *Review of Economic Studies* 49: 183–201.

Bourguignon, F. 1979. Decomposable income inequality measures. *Econometrica* 47: 901–920.

Chernoff, H. 1954. On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* 25: 573–578.

Cowell, F.A., and K. Kuga. 1981. Inequality measurement: An axiomatic approach. *European Economic Review* 13: 147–159.

Cox, D.R. 1961. Test of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on mathematical statistics and probability*, vol. 1. Berkeley: University of California Press.

De Jongh, B.H. 1952. *Egalisation, disparity and entropy*. Utrecht: A.W. Bruna en Zoons-Uitgevers Maatschapij.

Fisher, R.A. 1925. *Statistical methods for research workers*. London: Oliver & Boyd.

Foster, J.E. 1983. An axiomatic characterization of the Theil measure of income inequality. *Journal of Economic Theory* 31(1): 105–121.

Judge, G.G., W.E. Griffiths, R.C. Hill, and T.C. Lee. 1980. *The theory and practice of econometrics*. New York: Wiley.

Kolm, S-Ch. 1977. Multi-dimensional egalitarianism. *Quarterly Journal of Economics* 91: 1–13.

Kullback, S. 1959. *Information theory and statistics*. New York: Wiley.

Lisman, J.H.C. 1949. Econometrics and thermodynamics: A remark on Davis' theory of budgets. *Econometrica* 17: 59–62.

Maasoumi, E. 1986a. The measurement and decomposition of multi-dimensional inequality. *Econometrica* 54(5).

I

Maasoumi, E. 1986b. Unknown regression functions and information efficient functional forms: An interpretation. In *Innovations in quantitative economics*, ed. D. Slottje, Greenwich: JAI Press.
Pikler, A. 1951. Optimum allocation in econometrics and physics. *Welwirtschaftliches Archiv* 66: 97–132.
Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423, 623–56.
Shorrocks, A.F. 1980. The class of additively decomposable inequality measures. *Econometrica* 48: 613–625.
Stumpers, F.L.H.M. 1953. A bibliography of information theory; Communication theory–cybernetics. *IRE Transactions*, PGIT-2.
Theil, H. 1967. *Economics and information theory.* Amsterdam: North-Holland.
Theil, H. 1972. *Statistical decomposition analysis with applications in the social and administrative sciences*. Amsterdam: North-Holland.
Theil, H. 1980. The increased use of statistical concepts in economic analysis. In *Developments in statistics*, vol. 3, ed. P.R. Krishnaiah. New York: Academic Press.
Theil, H., and D. Fiebig. 1984. *Exploiting continuity: Maximum entropy estimation of continuous distributions*. Cambridge, MA: Ballinger.
Theil, H., and K. Laitinen. 1980. Singular moment matrices in applied econometrics. In *Multivariate analysis V*, ed. P.R. Krishnaiah. Amsterdam: North-Holland.
Wiener, N. 1948. *Cybernetics*. New York: John Wiley.

Infrastructure and Growth

César Calderón and Luis Servén

Abstract

An adequate supply of infrastructure services has long been considered essential for economic development by both academics and policymakers. This article reviews recent theoretical and empirical literature on the effects of infrastructure development on growth. The theoretical literature has employed a variety of analytical settings regarding the drivers of income growth and the degree to which infrastructure represents a public or a private good. In turn, the empirical literature has tested for the growth effects of infrastructure development using various econometric methodologies on time series and cross-section macro- and microeconomic data. However, the empirical tests face challenging issues of measurement, identification and heterogeneity. Overall, the literature finds positive effects of infrastructure development on income growth. Still, the precise mechanisms through which these effects accrue, and their full impact on welfare, remain relatively unexplored.

Keywords

Growth; Infrastructure; Public investment

JEL Classification

O40; D31

Introduction

An adequate supply of infrastructure services has long been considered essential for economic development by both academics and policymakers. The role of transport infrastructure, for instance, in fostering economic prosperity goes back to Adam Smith's *Wealth of Nations*, which listed 'the duty of erecting and maintaining certain public works' among the three core obligations of the sovereign.

Over the last quarter century, research has devoted considerable attention to the contribution of infrastructure development to the growth of productivity and aggregate income. A vast literature has explored a multitude of theoretical scenarios characterising the economic role of productive public services and their financing. In turn, a large empirical literature has examined the evidence on the growth impact of infrastructure development in a variety of cross-section, time-series and panel data settings. As discussed below, however, these empirical assessments are subject to a number of methodological caveats.

Infrastructure and Growth: Theory

Starting with the work of Aschauer (1989), a vast analytical and empirical literature has been concerned with the effects of infrastructure

development on income growth, productivity and welfare. Below is a summary view; more comprehensive accounts can be found in Irmen and Kuehnel (2009) and Romp and De Haan (2007).

Much of the relevant analytical literature examines the growth effects of productive public expenditure rather than just infrastructure. The two concepts are not necessarily equivalent. The government's involvement in productive activities is not limited to infrastructure sectors in many countries. Also, the public sector has traditionally played the leading role in the provision of infrastructure; however, private sector participation is growing in an increasing number of countries.

Following the seminal work of Arrow and Kurz (1970), the output impact of infrastructure has been modelled by including either the stock of infrastructure assets or the flow of infrastructure services as an additional input in the economy's aggregate production function, and further assuming that infrastructure is a gross complement for non-infrastructure inputs – labour and non-infrastructure capital. In this framework, an increase in the volume of infrastructure services raises output not only directly, but also indirectly, by 'crowding-in' other inputs owing to the accompanying rise in their marginal productivity. This indirect effect may take place instantaneously (for variable inputs in elastic supply) or over time (for fixed inputs such as human and non-infrastructure physical capital).

However, the expansion of infrastructure needs to be financed, and this represents a countervailing force: increasing taxation to finance public infrastructure crowds out the use of other inputs, which offsets partly or fully the crowding-in effect via productivity. This was highlighted by Barro (1990) in an endogenous growth framework in which the government's contribution to current output is captured by the flow of productive public expenditure (rather than the stock of public capital) financed through proportional income taxation. The welfare-maximising level of productive expenditure is shown to be the same as that which maximises the economy's growth rate, and it is achieved when the share of productive government expenditure in GDP (and hence the tax rate) equals the elasticity of aggregate output with respect to the same variable – what is often called the '*Barro rule*'. If productive expenditure exceeds this level, the additional distortionary taxation needed to finance it diverts non-infrastructure investment away to the point that income growth is reduced. These results apply in more general settings; for instance, Fisher and Turnovsky (1998) show that, in the Ramsey-type framework of Arrow and Kurz (1970) and with Cobb–Douglas technology, an increase in the stock of public infrastructure raises the private capital stock only if infrastructure spending is below the level defined by the Barro rule. With more general technologies, the elasticity of substitution between infrastructure and other capital also comes into play; see Eden and Kraay (2014).

Many of the theoretical contributions after Barro (1990) use an endogenous growth framework allowing infrastructure to impact the economy's long-run growth rate. However, in many cases (including Aschauer's (1989) pioneering empirical analysis), the focus is on the stock of infrastructure assets rather than the flow of infrastructure-related expenditure. The underlying logic is that, while the flow-based approach offers the important advantage of analytical tractability, the availability of infrastructure services (e.g. road transport) often relates more closely to the stock of infrastructure assets (e.g. the stock of public highways) than to the flow of expenditure on infrastructure-related activities (e.g. annual spending on road construction).

Following this logic, Futagami et al. (1993) extended Barro's (1990) model to include both public and private capital, with the rate of public investment as the government's key decision variable. This framework yields some new results. On the one hand, the economy displays nontrivial transitional dynamics. On the other, the growth-maximising level of public investment (as share of output) is still equal to the elasticity of output with respect to public capital; however, its welfare-maximising level is lower. Intuitively, public investment takes time to become productive, and this delay entails an additional sacrifice of current consumption for future consumption.

I

In reality, infrastructure provision requires both capital and recurrent expenditure – for example to build and maintain roads, respectively. Tsoukis and Miller (2003) and Ghosh and Roy (2004) examine how the above results are affected when the stock of public capital and the flow of non-investment spending are considered simultaneously. Overall, the earlier results stand: the welfare and growth-maximising levels of recurrent expenditure coincide, but they differ in the case of investment expenditure, for which growth maximisation implies public investment in excess of the welfare-maximising level.

Modelling infrastructure just like another input in production is a natural way to capture producers' direct use of electricity or transport services. But infrastructure may also enter the production function as a determinant of aggregate TFP, i.e. an 'unpaid factor' with spillover effects on the productivity of other inputs (Hulten and Schwab 2000). For example, Bougheas et al. (2000) and Agénor (2013) argue that transport and telecommunications services facilitate innovation and technological upgrading by reducing the fixed cost of producing new varieties of intermediate inputs. In a Romer-style framework, this raises output growth.

Aside from its role in the production function, another strand of the literature highlights the role of infrastructure in the accumulation of other inputs. For example, better transport networks may reduce the installation costs of new capital (Turnovsky 1996). Likewise, improved access to electricity may raise educational attainment and reduce the cost of human capital accumulation (Agénor 2011). In these cases, the growth-maximising output share of infrastructure spending is not given just by the elasticity of output with respect to infrastructure capital; one must also account for the output effect accruing through the accumulation of other inputs, and this tends to make the growth-maximising rate of infrastructure provision (as well as its welfare-maximising level) higher than when the latter effect is absent.

Contrary to what much of the literature assumes, few infrastructure services are pure public goods. In particular, congestion tends to make most services rival; think of road transportation, for example. Further, many services – such as power and telecommunications, or even toll roads – are excludable, and thus suitable for financing through user fees (and for private provision); see Ott and Turnovsky (2006) for a discussion of this issue.

The literature has considered two forms of infrastructure congestion. Under absolute congestion, services received by an individual user depend negatively on aggregate usage. Under relative congestion, they depend positively on the individual's usage relative to aggregate usage. Analytical details vary depending on the chosen option as well as the production technology (Barro and Sala-i-Martin 1992; Eicher and Turnovsky 2000). Nevertheless, some basic results from the models without congestion continue to hold. For example, in an endogenous growth setting with infrastructure modelled as a service flow, the welfare-maximising level of public infrastructure spending is still dictated by the Barro rule. If infrastructure is viewed instead as a stock, such a rule leads to excessive accumulation, just like in the absence of congestion (Turnovsky 1997). However, in an exogenous growth setting, the crowding-in effect of infrastructure on non-infrastructure capital tends to be diminished, or even reversed, especially when the financing is done through distortionary taxes (Fisher and Turnovsky 1998).

Another important feature of infrastructure is the presence of network effects, which can lead to strong nonlinearities in its marginal productivity. For example, road construction may have limited effects until the road network is minimally developed, at which point the marginal output contribution of additional roads may rise sharply. Once the entire network has been completed, however, additional road building is likely to have rapidly declining output effects (see Fernald 1999). Under appropriate conditions, these nonlinearities may lead to multiple equilibria, and to an enhanced role of infrastructure development policy: with a poor infrastructure endowment, the marginal productivity of infrastructure is low, and only a low-growth equilibrium may be attainable by the economy. However, a sufficient expansion of infrastructure networks would raise the

productivity of infrastructure and permit reaching the high-growth equilibrium (Agénor 2013).

Empirical Evidence

Few in academic or policy circles would dispute the view that infrastructure development fosters growth, but there is little consensus on the actual size of the effect and the factors that shape it. The empirical literature concerned with this issue took off following Aschauer (1989), who found that the stock of public infrastructure capital is a significant determinant of US aggregate TFP. However, his estimates of the elasticity of output with respect to infrastructure were implausibly large (around 0.40), owing to problems of econometric specification (see Gramlich 1994).

The massive empirical literature that followed focused on the impact of infrastructure on the level and growth rate of aggregate output or productivity, with numerous papers employing a large variety of data and empirical methodologies. Many authors estimated the elasticity of GDP with respect to infrastructure in an aggregate production function setting, using national or subnational data and time-series or panel techniques suitable for dealing with nonstationary variables and spillover effects. Early applications to panel data on US states found much smaller elasticities than those estimated by Aschauer (e.g. Holtz-Eakin 1994; Baltagi and Pinnoi 1995). Drawing from a large number of subsequent empirical studies using aggregate data, primarily from industrial countries, a meta-regression analysis of the elasticity of output with respect to public capital yields an average estimate of around 0.10, although the individual estimates from the underlying studies vary widely, from −1.73 to + 2.04 (Bom and Ligthart 2014).

These studies use monetary measures of public capital, constructed by accumulating investment flows. Alternatively, others employ physical measures of infrastructure assets encompassing multiple infrastructure sectors – sometimes aggregated into a synthetic indicator. Empirical studies using the latter approach on cross-country panel data typically report a significant GDP (or productivity) contribution of infrastructure; see Canning (1999), Calderón and Servén (2003) and Calderón et al. (2014). Regarding individual sectors, Röller and Waverman (2001) find a large output impact of telecommunications infrastructure in industrial countries while Fernald (1999) reports similar results for roads using US industry-level data.

In the Cobb–Douglas framework used by many empirical papers, it is not possible to assess the extent to which the effects of infrastructure reflect its TFPaugmenting role. This is the focus of relatively few studies. Hulten and Schwab (2000) use a growth decomposition approach to examine the contribution of public capital to manufacturing TFP growth across US states. They fail to find significant effects. Hulten et al. (2006) apply a similar approach to data from Indian states, using physical indicators of infrastructure assets in transport and power, and find that infrastructure development accounted for almost half of the observed TFP growth. Duggal et al. (1999) allow for nonlinear production technologies. Using aggregate US data, they find that public infrastructure capital is an important determinant of TFP. Duggal et al. (2007) extend the framework to include also privately supplied IT infrastructure, which contributes to production both as a standard input and as a driver of TFP. Both types of infrastructure are found to have a significant positive effect on productivity.

A related line of research, pioneered by Berndt and Hansson (1991), takes a dual approach and focuses on the estimation of cost and/or profit functions augmented by either infrastructure or public capital stock measures. The empirical finding in most cases is that infrastructure reduces production costs or increases profits – see Demetriades and Mamuneas (2000) on OECD cross-country data, and Cohen and Morrison Paul (2004) on US state data.

A different strand of literature evaluates the long-term growth impact of infrastructure, typically using a reduced-form growth-regression framework relating long-run growth to suitable indicators of infrastructure, public capital or public investment, often in conjunction with standard control variables from the empirical growth literature. Measures of infrastructure and conditioning

variables differ across studies, so they are not easy to compare. However, those papers using monetary measures of public capital stocks or public investment yield mixed results – for example Holtz-Eakin and Schwartz (1995) and Crihfield and Panggabean (1995) find no significant growth effects of infrastructure across US states and metropolitan areas. In turn, Easterly and Rebelo (1993) find that public investment in transport and communications significantly raises growth across countries. Devarajan et al. (1996) find a negative relationship between the share of infrastructure in total public expenditure and economic growth in panel data for developing countries, while Gupta et al. (2005) find the opposite result in a different cross-country panel data set.

In contrast, growth regressions using physical indicators of infrastructure stocks almost invariably find significant growth effects. In many cases, they use the number of telephone lines proxy for infrastructure (e.g. Easterly 2001). In others, they use synthetic indicators capturing physical stocks in multiple infrastructure sectors – transport, power and telecommunications. Sánchez-Robles (1998) and Calderón and Servén (2004, 2010a, b) find that these summary measures are positively and robustly related to per capita GDP growth in panel datasets combining industrial and developing countries. The magnitude of the effects is substantial: a 1% increase in physical infrastructure stocks, given other variables, temporarily raises GDP growth by as much as 1–2 percentage points, although the growth acceleration gradually tapers off as the economy approaches its long-run per capita income.

The literature cited so far takes a country-level perspective. However, there are also studies that examine the effects of infrastructure development for income growth at a more disaggregated level. For example, Rud (2012) investigates the impact of electricity provision on manufacturing output across Indian states. Electricity provision is not exogenously assigned, and to deal with this problem the study takes advantage of the introduction of a new irrigation-intensive agricultural technology, viewed as a natural experiment. Adoption of new varieties of high-yield seeds required, among other things, timely irrigation, for which electric pumps were used. Thus the initial availability of groundwater across states is employed to control for the endogeneity of the expansion of the electricity network. The evidence shows that, on average, a one standard deviation increase in the measure of electrification is associated with a 14% expansion in state manufacturing output.

In turn, Datta (2012) examines the consequences of a major road improvement program in India – the Golden Quadrilateral Program (GQP) – for the performance of firms. The location of each individual firm relative to the upgraded highway provides firm-specific exogenous variation in the degree to which the quality of the roads improved as a result of the GQP. The study finds that firms located on the GQP-improved highways significantly enhanced their inventory management and reduced their input costs by switching suppliers.

The bulk of the empirical literature summarised here focuses on measuring the output (or productivity) gains from infrastructure assets. Less attention has been paid to the cost of acquiring and operating these assets. Yet comparison of (social) marginal costs and benefits is necessary to determine whether infrastructure is under- or over-provided. To some extent, this is implicitly done in the reduced-form growth regressions mentioned earlier, given that their estimates of the impact of infrastructure allow for the adjustment of other production inputs as well as the changes in fiscal parameters required to accommodate infrastructure shocks. However, the issue is ignored in most other studies.

Among the exceptions are Canning and Pedroni (2008), who use a simple empirical model in the spirit of Barro (1990) to compare physical infrastructure stocks with their growth-maximising levels in a panel of countries. Their finding is that infrastructure is under-provided in some countries and over-provided in others, and the verdict shows no clear correlation with countries' per capita income. On average, the level of infrastructure is 'just about right' from the point of view of growth maximisation, so there is no evidence of a generalised infrastructure shortage. Using a similar framework, Kamps (2005) likewise concludes that there is no shortage of public

capital in EU countries. In turn, Eden and Kraay (2014) assess public capital shortages in low-income countries, using a Ramsey-type framework that highlights the degree of substitutability between public and private capital. Their estimate of the marginal return on public capital exceeds the user cost, given by the rate of depreciation plus the world real interest rate (thus implicitly assuming nondistortionary taxation). They conclude that, on average, public capital is under-provided in their sample countries.

Limitations of the Empirical Literature

Three major concerns arise from the empirical literature on the development impact of infrastructure: measurement, identification and heterogeneity. Take *measurement* first. Infrastructure is a multi-dimensional concept, comprising services that range from transport to clean water. However, many studies take a single indicator (most often telephone density) to proxy for 'infrastructure'. Omitting other indicators of infrastructure where they are relevant – e.g. in growth empirics – leads to invalid inferences. However, simultaneous consideration of multiple types of infrastructure assets in econometric estimation will often lead to imprecise estimates. This motivates the use of synthetic infrastructure indices – see Sánchez-Robles (1998) and Calderón and Servén (2004, 2010a, b).

Furthermore, measures of infrastructure based on spending flows – typically, public investment or its accumulation via perpetual inventory into public capital stocks – pose their own problems. Public investment and public capital are likely to be poor proxies for infrastructure accumulation if private participation in infrastructure provision is significant, as has become the case in many countries, or if the public investment is partly allocated to non-infrastructure industrial and commercial activities of the public sector. And even aside from these caveats, the link between observed public capital expenditure and the accumulation of infrastructure assets or the provision of services – which is what really matters for growth and equity – may be weak or nonexistent, owing to inefficiencies in public procurement and outright corruption (Pritchett 2000; Keefer and Knack 2007). Furthermore, investment may not translate into commensurate increases in the supply of infrastructure services because of inefficiencies in the selection and implementation of projects or the absence of high-quality projects in the pipeline (Kilby 2013). These limitations do not apply to physical measures of infrastructure, which may be the reason why studies based on them are more conclusive than those based on monetary measures of infrastructure.

Identification remains a thorny issue. The impact of infrastructure supply on growth that empirical studies aim to estimate may be confounded with increased demand for infrastructure services prompted by rising levels of income (Canning and Pedroni 2008).

There is no easy solution for this problem. In theory, one could base inference on the estimation of a full structural model. However, that approach poses difficult specification choices and challenging data requirements. Esfahani and Ramírez (2003) and Cadot et al. (2006) present two-equation models that jointly describe the aggregate production function and the accumulation of infrastructure. The former paper highlights the role of institutional factors for accumulation decisions in a cross-country setting, while the latter puts emphasis on the political economy of investing in transport routes across French provinces. Both papers find that the contribution of infrastructure services to GDP more than exceeds the cost of providing them. Interestingly, Cadot et al. (2006) find that the estimated elasticity of output with respect to infrastructure (around 0.08–0.09) remains invariant regardless of whether one accounts for the likely endogeneity.

Another option recently used by Calderón et al. (2014) in a panel time-series setting is to establish the existence of a single long-run relation between infrastructure, aggregate output and other production inputs, which can then be interpreted as the economy's aggregate production function. If infrastructure and the other productive inputs do not react systematically to temporary deviations from the long-run relation, its parameters can be estimated by conventional single-equation methods (even if the parameters

I

characterising the short-run dynamics cannot). Formal exogeneity tests confirm that this is the case, and the estimation places the long-run elasticity of output with respect to a synthetic index of infrastructure in the range 0.08–0.10.

A third alternative is to use an instrumental variable approach, ideally featuring external instruments for infrastructure. In this vein, Calderón and Servén (2003, 2004) employ demographic variables as instruments, alone or in combination with internal instruments, in a GMM panel framework. Röller and Waverman (2001) follow a similar approach.

Lastly, *heterogeneity* is also a major concern. The contribution of infrastructure to income growth may vary across countries and over time for multiple reasons – starting with the heterogeneous quality of infrastructure assets and services themselves. However, few studies take into account the quality dimension, in large part due to data limitations. Hulten (1996) finds that differences in the effective use of infrastructure resources explain one-quarter of the growth differential between Africa and East Asia, and more than 40% of the growth differential between lowand high-growth countries. Rioja (2003) likewise finds that poor infrastructure quality imposes large output and welfare costs across Latin American countries. Calderón and Servén (2004, 2010a, b) and Seneviratne and Sun (2013) find significant growth effects of a synthetic indicator of infrastructure quality in an empirical framework including also the quantity of infrastructure.

Aside from asset quality, variation across space and time in the effects of infrastructure could arise from many other sources, such as network effects that create nonlinearities in the output contribution of infrastructure and institutional factors that constrain the efficient use of infrastructure assets. However, assessments of heterogeneity are not abundant in the empirical literature. As an exception, Calderón et al. (2014) test for parameter heterogeneity in a large cross-country sample, using an infrastructure-augmented production function framework. Their tests consider heterogeneity across countries both of unrestricted form – affecting the parameters of infrastructure or any other input – as well as heterogeneity in the effects of infrastructure related to specific country features. These include the level of per capita GDP, the extent of infrastructure development and the size of population – to capture network and congestion effects, respectively. All these tests fail to reject homogeneity. The implication is that, other things equal, the percentage increase in real GDP (or its growth rate) that results from a given percentage increase in the availability of infrastructure does not vary much across countries. In the paper's setting, this means that the marginal productivity of infrastructure is higher, other things being equal, where the (relative) stock of infrastructure is lower.

Final Remarks

In spite of the above caveats, the balance of empirical research does reveal a positive contribution of infrastructure development to aggregate income. In itself, this just confirms that the marginal productivity of infrastructure capital is positive. But there has been also some convergence in quantitative estimates of its magnitude, to levels generally much lower than those found in the earlier macro literature. Still, the precise mechanisms at work remain understudied – including, for example, the role of infrastructure quality, the extent of crowding-in effects and the significance of the TFP channel of transmission. Furthermore, in contrast with the effort devoted to quantify the output impact of infrastructure, research has paid much less attention to the costs of infrastructure development. As a consequence, there are few empirical results regarding the extent to which different infrastructure services may be over- or under-provided across countries or regions. In this context, one key ingredient in need of more attention is the fragile link between infrastructure expenditures and the accumulation of infrastructure assets or the provision of services, and especially how such a link is shaped by institutional and political economy factors.

See Also

► Economic Growth
► Extreme Poverty

► Growth and Inequality (Macro Perspectives)
► Infrastructure and Inequality
► Public Infrastructure

Acknowledgments We thank Jon Temple for useful comments and suggestions. The views expressed here are ours only and do not necessarily reflect those of the World Bank, its Executive Directors or the countries they represent.

Bibliography

Agénor, P.R. 2011. Schooling and public capital in a model of endogenous growth. *Economica* 78: 108–132.

Agénor, P.R. 2013. *Public capital, growth and welfare*. Princeton: Princeton University Press.

Arrow, K., and M. Kurz. 1970. *Public investment, the rate of return and optimal fiscal policy*. Baltimore: Johns Hopkins University.

Aschauer, D. 1989. Is public expenditure productive? *Journal of Monetary Economics* 23: 177–200.

Baltagi, B., and N. Pinnoi. 1995. Public capital and state productivity growth. *Empirical Economics* 20: 351–359.

Barro, R.J. 1990. Government spending in a simple model of exogenous growth. *Journal of Political Economy* 98: 103–125.

Barro, R., and X. Sala-i-Martin. 1992. Public finance in models of economic growth. *Review of Economic Studies* 59: 645–661.

Berndt, E.R., and B. Hansson. 1991. *Measuring the contribution of public infrastructure capital in Sweden*. NBER Working Paper 3842.

Bom, P., and J. Ligthart. 2014. What have we learned from three decades of research on the productivity of public capital? *Journal of Economic Surveys*, forthcoming.

Bougheas, S., P. Demetriades, and T. Mamuneas. 2000. Infrastructure, specialization and economic growth. *Canadian Journal of Economics* 33: 506–522.

Cadot, O., L. Röller, and A. Stephan. 2006. Contribution to productivity or pork barrel? The two faces of infrastructure investment. *Journal of Public Economics* 90: 1133–1153.

Calderón, C., and L. Servén. 2003. The output cost of Latin America's infrastructure gap. In *The limits of stabilization: Infrastructure, public deficits, and growth in Latin America*, ed. W. Easterly and L. Servén, 95–118. Stanford University Press and the World Bank.

Calderón, C., and L. Servén. 2004. *The effects of infrastructure development on growth and income distribution*. World Bank Policy Research Working Paper 3400.

Calderón, C., and L. Servén. 2010a. Infrastructure and economic development in Sub-Saharan Africa. *Journal of African Economies* 19(S1): 13–87.

Calderón, C., and L. Servén. 2010b. Infrastructure in Latin America. In *The Oxford handbook of Latin American economies*, ed. J. Ocampo and J. Ros. Oxford: Oxford University Press.

Calderón, C., E. Moral-Benito, and L. Servén. 2014. Is infrastructure capital productive? A panel heterogeneous approach. *Journal of Applied Econometrics*, forthcoming.

Canning, D. 1999. *The contribution of infrastructure to aggregate output*. World Bank Policy Research Working Paper 2246.

Canning, D., and P. Pedroni. 2008. Infrastructure and long-run economic growth. *The Manchester School* 76(5): 504–527.

Cohen, J.P., and C.J. Morrison Paul. 2004. Public infrastructure investment, interstate spatial spillovers, and manufacturing costs. *Review of Economics and Statistics* 86(2): 551–560.

Crihfield, J., and M. Panggabean. 1995. Is public infrastructure productive? A metropolitan perspective using new capital stock estimates. *Regional Science and Urban Economics* 25: 607–630.

Datta, S. 2012. The impact of improved highways on Indian firms. *Journal of Development Economics* 99: 46–57.

Demetriades, P., and T. Mamuneas. 2000. Intertemporal output and employment effects of public infrastructure capital: Evidence from 12 OECD economies. *The Economic Journal* 110: 687–712.

Devarajan, S., V. Swaroop, and H.F. Zou. 1996. The composition of public expenditure and economic growth. *Journal of Monetary Economics* 37: 313–344.

Duggal, V., C. Saltzman, and L. Klein. 1999. Infrastructure and productivity: A nonlinear approach. *Journal of Econometrics* 92: 47–74.

Duggal, V., C. Saltzman, and L. Klein. 2007. Infrastructure and productivity: An extension to private infrastructure and IT productivity. *Journal of Econometrics* 140: 485–502.

Easterly, W. 2001. The lost decades: Explaining developing countries' stagnation in spite of policy reform 1980–1998. *Journal of Economic Growth* 6: 135–157.

Easterly, W., and S. Rebelo. 1993. Fiscal policy and economic growth: An empirical investigation. *Journal of Monetary Economics* 32: 417–458.

Eden, M., and A. Kraay. 2014. *'Crowding in' and the returns to government investment in low-income countries*. World Bank Policy Research Working Paper 6781.

Eicher, T., and S. Turnovsky. 2000. Scale, congestion and growth. *Economica* 67: 325–346.

Esfahani, H., and M. Ramírez. 2003. Institutions, infrastructure and economic growth. *Journal of Development Economics* 70(2): 443–477.

Fernald, J.G. 1999. Roads to prosperity? Assessing the link between public capital and productivity. *American Economic Review* 89: 619–638.

Fisher, W.H., and S.J. Turnovsky. 1998. Public investment, congestion, and private capital accumulation. *Economic Journal* 108: 339–413.

Futagami, K., Y. Morita, and A. Shibata. 1993. Dynamic analysis of an endogenous growth model with public

capital. *Scandinavian Journal of Economics* 95: 607–625.
Ghosh, S., and U. Roy. 2004. Fiscal policy, long-run growth, and welfare in a stock–flow model of public goods. *Canadian Journal of Economics* 37: 742–756.
Gramlich, E.M. 1994. Infrastructure investment: A review essay. *Journal of Economic Literature* 32: 1176–1196.
Gupta, S., B. Clements, E. Baldacci, and C. Mulas-Granados. 2005. Fiscal policy, expenditure composition, and growth in low-income countries. *Journal of International Money and Finance* 24: 441–463.
Holtz-Eakin, D. 1994. Public sector capital and the productivity puzzle. *Review of Economics and Statistics* 76: 12–21.
Holtz-Eakin, D., and A. Schwartz. 1995. Infrastructure in a structural model of economic growth. *Regional Science and Urban Economics* 25: 131–151.
Hulten, C. 1996. *Infrastructure capital and economic growth: How well you use it may be more important than how much you have*. NBER Working Paper 5847.
Hulten, C., and R. Schwab. 2000. Does infrastructure investment increase the productivity of manufacturing industry in the U.S.? In *Econometrics and the cost of capital: Essays in honor of Dale Jorgenson*, ed. L. Lau. Cambridge, MA: MIT Press.
Hulten, C., E. Bennathan, and S. Srinivasan. 2006. Infrastructure, externalities, and economic development: A study of Indian manufacturing industry. *World Bank Economic Review* 20: 291–308.
Irmen, A., and J. Kuehnel. 2009. Productive government expenditure and economic growth. *Journal of Economic Surveys* 23(4): 692–733.
Kamps, C. 2005. Is there a lack of public capital in the European Union? *EIB Papers* 10: 72–93.
Keefer, P., and S. Knack. 2007. Boondoggles, rent-seeking and political checks and balances: Public investment under unaccountable governments. *Review of Economics and Statistics* 89: 566–572.
Kilby, C. 2013. The political economy of project preparation: An empirical analysis of World Bank projects. *Journal of Development Economics* 105: 211–225.
Ott, I., and S. Turnovsky. 2006. Excludable and non-excludable public inputs: Consequences for economic growth. *Economica* 73: 725–748.
Pritchett, L. 2000. The tyranny of concepts: CUDIE (Cumulated, Depreciated, Investment Effort) is not capital. *Journal of Economic Growth* 5(4): 361–384.
Rioja, F. 2003. The penalties of inefficient infrastructure. *Review of Development Economics* 7: 127–137.
Röller, L.-H., and L. Waverman. 2001. Telecommunications infrastructure and economic development: A simultaneous approach. *American Economic Review* 91: 909–923.
Romp, W., and J. DeHaan. 2007. Public capital and economic growth: A critical survey. *Perspektiven der Wirtschaftspolitik* 8(s1): 6–52.
Rud, J.P. 2012. Electricity provision and industrial development: Evidence from India. *Journal of Development Economics* 97: 352–367.
Sánchez-Robles, B. 1998. Infrastructure investment and growth: Some empirical evidence. *Contemporary Economic Policy* 16: 98–108.
Seneviratne, D., and Y. Sun. 2013. *Infrastructure and income distribution in ASEAN-5: What are the links?* IMF Working Paper 13/41.
Tsoukis, C., and N. Miller. 2003. Public services and endogenous growth. *Journal of Policy Modeling* 25: 297–307.
Turnovsky, S. 1996. Fiscal policy, adjustment costs, and endogenous growth. *Oxford Economic Papers* 48: 361–381.
Turnovsky, S. 1997. Fiscal policy in a growing economy with public capital. *Macroeconomic Dynamics* 1: 615–639.

Infrastructure and Inequality

César Calderón and Luis Servén

Abstract

An adequate supply of infrastructure services has long been considered essential for economic development by both academics and policymakers. This article reviews recent theoretical and empirical literature on the effects of infrastructure development on income distribution. The theoretical literature has employed a variety of analytical settings regarding the dynamics of income distribution, the extent of market distortions – notably in capital markets – and the externalities surrounding infrastructure services. In turn, the empirical literature has tested for the distributive effects of infrastructure development using multiple approaches, from cross-country aggregate data to micro-level studies of specific infrastructure interventions. However, these empirical tests face challenging issues of measurement, identification and heterogeneity. Overall, the empirical evidence suggests that infrastructure development may have a positive effect on distributive equity. Still, little is known about the likely magnitude of such an effect and the precise mechanisms through which it may accrue.

Keywords
Inequality; Infrastructure; Poverty; Public investment

JEL Classifications
D31; H54

Introduction

Infrastructure has long been regarded as an essential ingredient for economic development. Over the last quarter century, a large body of analytical and empirical research has examined the contribution of infrastructure development to the growth of productivity and aggregate income (see the article on '▶ Infrastructure and Growth'). More recently, however, there has been growing recognition that, in addition to its impact on average productivity and income, infrastructure can also affect income inequality, and this issue has attracted increasing attention from theoretical and empirical research.

Conceptually, there are good reasons why infrastructure development could have a differential positive effect on the incomes of the poor, over and above its effect on aggregate income. Infrastructure facilitates the poor's access to productive opportunities – for example, by helping connect lower-income segments of the population to markets for their inputs and outputs, so that their incomes may rise more than the average, as may the value of their assets (land or human capital). Infrastructure can also help improve health and education outcomes for the poor, thus enhancing their human capital. More broadly, access to and use of infrastructure services – telecommunications, electricity, roads, safe water and sanitation – play a key role in the integration of individuals and households into social and economic life (World Bank 2003).

The theoretical literature on the linkages between infrastructure and inequality is not as vast as that on infrastructure and growth. It has examined the distributional effects of infrastructure development under various assumptions about income distribution dynamics, economic distortions – notably in capital markets – and infrastructure-driven externalities. Empirical research has likewise employed a variety of approaches, from cross-country and time-series regressions using macroeconomic data, to micro-level studies of the effects of specific infrastructure interventions on the incomes of the poor, especially in rural areas. However, as discussed below, these empirical assessments are not free of methodological caveats.

Theory

Attempts to model the relationship between public investment and inequality are grounded in the literature on wealth distribution dynamics in the presence of capital market imperfections – see Banerjee and Newman (1993), Galor and Zeira (1993) and Piketty (1997). In these models, wealth redistribution towards the poor or the middle class can improve productive efficiency (Aghion and Bolton 1992, 1997). Enhanced availability of productive services – such as education, health and infrastructure – to the general population may not only improve efficiency, but also help reduce inequality. In this vein, Ferreira (1995) builds a model with private–public capital complementarity in an environment with capital market imperfections. The government participates in the production of certain goods and services in which it has a comparative advantage (e.g. infrastructure, education and health), and only higher-income individuals can afford to purchase private alternatives to public services. In this context, expanding public infrastructure services reduces the inequality of opportunity among entrepreneurs, increases the return on investment and raises entrepreneurial activity among the less-favoured segments of society.

Building on this framework, more recent contributions model the joint dynamics of public investment, growth and inequality in a general equilibrium setting with heterogeneous agents that differ in their initial endowments of private capital. In these models, a pure public good or service (e.g. infrastructure) interacts with private capital in the production of other goods.

Getachew (2010) presents a two-sector growth model with capital market imperfections in which public capital not only contributes to the production of goods, but also promotes the accumulation of private (human) capital. Like in earlier models, income inequality hinders growth. Increased provision of productive public services not only raises aggregate growth, but can also influence the distribution of income (and thereby exert a further indirect impact on growth) if the services accrue heterogeneously across individual households. Specifically, greater provision of public infrastructure benefits the poor more than proportionally because of their lesser access to private substitutes.

Chatterjee and Turnovsky (2012) likewise examine the dual role of public capital as growth engine and determinant of inequality. In their setting, public capital affects both productivity and labour–leisure choices. Greater public investment raises factor incomes through the productivity channel, while also affecting relative factor returns and the distribution of income and welfare through the labour–leisure choice. However, the mode of financing public investment matters for factor income shares and income inequality. Numerical simulation of the model shows that any distributional gains may be only temporary if public investment is financed through non-distortionary taxes. On the other hand, income distribution improves in both the short and long run when public investment is financed by levies on capital.

Another dimension of income inequality that may be affected by public infrastructure development is the skill premium. It is examined by Pi and Zhou (2012) using a static multi-sectoral model with skilled and unskilled labour, in which public infrastructure is an input in the production of the different goods. A greater supply of public infrastructure raises the marginal productivity of both skilled and unskilled labour – and hence their respective remuneration. The effect on the skill premium depends on factor intensities: if the sector using unskilled labour is relatively more intensive in public infrastructure services, there will be an outflow of capital from the skilled to the unskilled sector. Hence the wage rate of skilled labour will decline and that of unskilled labour will increase. This reduces skilled–unskilled wage inequality. Of course, the effect is the opposite if the sector using skilled labour is more intensive in the use of the publicly provided infrastructure input.

The literature has devoted particular attention to the distributive impact of opening up infrastructure provision to private sector initiative. The impact may accrue through changes in employment, in the composition of public spending, and in the access and affordability of infrastructure services for the poor (Estache et al. 2000). Employment effects are particularly controversial, as former public enterprises acquired by private providers often become profitable by downsizing (Estache et al. 2002). In turn, the distributive impact of downsizing depends on the proportion of lower-income workers in the infrastructure labour force, and on the monetary compensation to laid-off workers. In addition, if the investment by newly reformed providers of infrastructure promotes growth and new jobs, downsizing in the public infrastructure sector may be offset by job creation in other sectors (Benitez et al. 2003).

Aside from employment effects, private sector participation in infrastructure also affects public revenues and expenditures. Subsidies to the provision of infrastructure services may be eliminated, and revenues from privatisation may be generated. What happens with inequality depends also on what is done with the increased fiscal resources. If they are devoted to improving the efficiency and coverage of public (infrastructure and/or non-infrastructure) service provision, income inequality may decline (Estache et al. 2000).

Finally, infrastructure reform may lead to price and/or supply responses that reduce the access and affordability of services for the poor. For example, removing subsidies may lead to higher post-reform prices, and new private providers may charge higher connection fees than government-owned providers or be reluctant to reach poorer areas (Estache et al. 2002). As a result, infrastructure services may become unaffordable to lower-income groups. The likelihood of this outcome depends on the overall design of the reforms. In practice, however, there are numerous episodes in

which access by the poor improved after reforms involving private participation.

Empirical Evidence

The empirical literature on infrastructure and inequality broadly follows two main strands. One is concerned with the effects of infrastructure stocks and/or service flows on standard inequality indicators. It includes the majority of the studies using macroeconomic data. The other examines the effects of specific infrastructure interventions, usually focusing on the income of poor households or backward geographic areas.

A few studies have directly examined the inequality impact of infrastructure at the aggregate level, by regressing Gini coefficients and similar inequality measures on indicators of infrastructure development in a cross-country panel data setting. Among them, López (2004) proxies infrastructure development by fixed telephone density, while Calderón and Chong (2004) consider the quantity and quality of different infrastructure sectors (telecommunications, energy, roads and railways), both separately and jointly, using a qualitative summary indicator in the spirit of Hulten (1996). In turn, Calderón and Servén (2004, 2010a, b) employ synthetic indices of infrastructure quantity and quality combining multiple infrastructure sectors, built through a principal components procedure. These papers find that, *ceteris paribus*, income inequality is negatively related to their respective measures of infrastructure development. In a similar setting, Seneviratne and Sun (2013) reach the same result for ASEAN countries, but they also find that public investment does not bear any significant relation to inequality. This suggests that public investment data offer a poor proxy for infrastructure development.

The literature also advances the testable hypothesis that increased access to infrastructure services should help raise the income and the value of the assets of the poor. However, the availability of information on access to infrastructure services varies dramatically across countries and infrastructure sectors. For telecommunications, water and sanitation, existing data provide fairly good coverage across countries and over time. For power and transport availability is sparse, especially in the time dimension. Subject to these constraints, Calderón and Servén (2010b) find a negative correlation across countries between measures of access to multiple infrastructure services and standard indicators of inequality – although it is not obvious to what extent this result may reflect a causal relation.

At the microeconomic level, another strand of literature examines the poverty effects of infrastructure interventions using matching techniques that combine samples of beneficiaries with samples drawn from regular household surveys (taken as control group). These studies usually evaluate the impact on income of a particular intervention affecting a given group of households or a specific geographic area.

Some studies of this type find that physical infrastructure in roads and communications facilitates spatial access and information flows, raising labour mobility, boosting rural non-farm economies and reducing the incidence of poverty in some geographic areas (Jalan and Ravallion 2003; Zhu and Luo 2006; Reardon et al. 2007). They also show that public infrastructure provides a boost for local community and market development. For instance, rehabilitating rural roads in Bangladesh raised non-agricultural wage employment in targeted households and fostered markets that have become increasingly diversified across sectors – with the largest impact on households in the second-lowest quartile of the income distribution being the most mobile in changing activity from agriculture to non-farm work (Khandker and Koolwal 2007, 2010). This type of intervention has also proved successful in Vietnam by increasing workers' wages and developing local markets in poor communities (Mu and van de Walle 2007).

Granting access to new and improved roads in rural areas has also expanded the set of opportunities in non-agricultural activities in Peru (Escobal and Ponce 2008) and in non-farm activities among women in Georgia (Lokshin and Yemtsov 2005). There is also evidence from large emerging markets such as China and India. For example, public

investment in rural roads and electrification has contributed to rapid growth in agricultural production across Chinese regions. However, the impact on poverty and inequality was boosted when infrastructure expansion was accompanied by public investment in education, science and technology (Fan and Zhang 2004; Zhang and Fan 2004). On the other hand, an expansion of regional infrastructure facilities (e.g. power and roads) in certain regions and districts of India was found to have improved average living standards and lowered the share of people living below the poverty line, even when infrastructure investment was accompanied by divestitures in education and health (Majumder 2012).

Recent literature examines the impact of electrification programs on rural areas in developing countries. Dinkelman (2011) evaluates the effect of the massive roll-out of the electricity grid in rural South Africa on employment – and, most notably, female employment. This roll-out, started in 1995, targeted low-capacity household use in rural areas rather than industrial users. The study employs the land gradients of the communities to adjust for the endogenous location of projects. The main finding is that electrification leads to rising female employment on both the extensive and the intensive margins. For instance, women work nearly 9 hours more per week in districts that experienced an average increase in electrification. This occurs as households with access to electricity replace wood burning at home with electricity for cooking and lighting, which frees up female time from home to market work. It also provides new opportunities to produce home-based goods and services for the market, through either self-employment or micro-enterprises.

One particular intervention found to have significant distributional effects is the construction of large irrigation dams. Duflo and Pande (2007) find that the benefits of building a dam on irrigated areas, in terms of agricultural production and rural poverty, accrue to the districts located downstream from the dam – as opposed to those districts were the dam is built. Furthermore, downstream districts can use the dam as insurance against rainfall shocks while agricultural production in districts where the dam is built is more vulnerable. In sum, rural poverty falls in districts located downstream, but this decline is smaller in magnitude than the increase in districts where the dam was built.

Other empirical studies shed light on the theoretical claim that improved access to infrastructure services can raise the income of the poor through its impact on human capital – specifically, education and health outcomes. Better transportation systems and safer roads help raise school attendance (Brenneman and Kerf 2002), while improved access to electricity allows more time for study and the use of computers (Leipziger et al. 2003). Cross-country research shows that enhanced access and use of basic infrastructure services reduces rates of child and maternal mortality. Likewise, Jalan and Ravallion (2003) find that the prevalence and duration of diarrhoea in children under five in rural India is lower among households with piped water, although the impact on the poor is amplified if public investment in water and sanitation is accompanied by other interventions in education and incomepoverty reduction. Analogous benefits resulted in Argentina when privatisation expanded access to water and sanitation by the poor – child mortality fell by 8% (Galiani et al. 2005).

Recent evidence shows that the benefits to the poor of improved access to water may go beyond the conventional health effects. Better access reduces time devoted to water collection, thereby freeing up time for additional leisure or production. It reduces important sources of stress and tension within the household and/or community. Moreover, it provides women greater mobility and the opportunity to socialise and improve their well-being. Overall, welfare gains may result even in the absence of income or health gains; see Devoto et al. (2011) for evidence from the city of Tangiers.

Finally, evidence from Latin America shows that privatisation of infrastructure sectors often benefited the poorest groups by granting them access to services. For instance, Chisari et al. (1999) and Navajas (2000) find that the privatisation of infrastructure services in Argentina hurt the middle class relatively more than the rest of the income groups through the elimination of existing subsidies. However, it benefited the

poor by improving their access to services. Estache et al. (2000) show that the less-favoured segments of the population in Latin America had very limited (or no) access to many utility services, and thus did not benefit from their expansion prior to the privatisation. However, the extent of the benefits from privatisation reaped by the poorest differed across sectors. In many countries, the rapid expansion of mobile telephone networks led to increased access to a wide array of service suppliers. The power sector, on the other hand, moved at a slower pace post-privatisation, and reforms often failed to provide low-cost solutions to remote households in rural areas (Foster et al. 2001). More broadly, an encompassing review of Latin America's experience offers several examples of improved access to infrastructure post-privatisation (World Bank 2003). For instance, improved access to electricity, water and telephones for poorer groups lifted their incomes in Guatemala. The expansion of infrastructure services to rural areas in El Salvador reduced the time required to reach markets, which created significant gains for poorer groups. Lastly, improving road quality had an important impact on non-agricultural income for the rural population in Peru, most notably arising from wage employment.

Limitations of the Empirical Literature

Empirical assessments of the distributional effects of infrastructure development face methodological challenges concerning measurement, identification and heterogeneity similar to those that affect studies of the growth impact of infrastructure. Since these are described at length under 'Infrastructure and growth', the discussion below highlights only the particular forms that these methodological issues take when assessing the distributive consequences of infrastructure development.

Regarding *measurement*, perhaps the key obstacle is the lack of systematic information on access to, and affordability of, infrastructure services for different percentiles of the income distribution, whether over time or across countries. This makes it very hard to reach robust conclusions regarding the consequences of infrastructure development for the equality of opportunities and incomes across households. Researchers have resorted to aggregate data on access – that is, without a breakdown across income percentiles – under the implicit assumption that changes in access at the margin affect primarily the poorer segments of the population, but this may not always be the case. More fundamentally, even aggregate access figures are available for only a limited number of countries, in most cases without any geographic disaggregation, and only for recent years.

Identification remains a challenging problem. While infrastructure may help reduce inequality, at the same time inequality may hamper the provision of infrastructure services to the poor. The reason is that more unequal societies devote fewer resources to the provision of public goods, including infrastructure (Alesina et al. 1999). Cross-country (or cross-region) studies that fail to account for this and similar forms of simultaneity are likely to overstate the effects of infrastructure development on equity. More broadly, unobserved factors affecting both distributional outcomes and infrastructure accumulation may likewise lead to biases.

Recent microeconomic studies of the impact of specific infrastructure interventions have used randomised control trials (RCTs) to establish causality. This approach enables researchers to assess whether any changes observed in the target population are due to the public infrastructure program, exogenous factors, or unmeasured individual effects. RCTs isolate the impact of interventions by randomly assigning individuals to treatment and control groups. In this vein, several studies have examined the impact of improved water and sanitation on health outcomes (Capuno et al. 2011; Andrés et al. 2014). A limitation of this approach, however, is that the findings may depend on the specific context and time frame in which the experiment is conducted, so evidence of a successful policy intervention might not be relevant to other cities, regions or countries.

Lastly, *heterogeneity* is also a relevant issue. The effects of infrastructure on inequality may vary across locations and over time for multiple

reasons. One is the heterogeneous quality of infrastructure assets and services themselves, difficult to account for due to data limitations. Attempts to bring infrastructure quality into the analysis, along with quantity, include Calderón and Servén (2004) and Seneviratne and Sun (2013). However, variation in the measured effects of infrastructure development could arise from many other sources. This is a particular relevant concern given the major role that studies of particular infrastructure interventions play in this empirical literature, because their findings may reflect a host of unmeasured (or hard-to-replicate) factors specific to the intervention under consideration.

Final Remarks

Overall, available empirical research offers some suggestive evidence that infrastructure development has equity-enhancing effects. The analytical literature has proposed a number of specific mechanisms through which such effects might accrue, but evidence on their actual relevance is, in most cases, still incomplete. Data limitations are largely responsible for this. Infrastructure development should affect poorer households primarily by improving their access to affordable services. However, the limited information available on access and affordability for households at different percentiles of the income distribution represents a major obstacle to progress in establishing unambiguously the consequences of infrastructure development for inequality and, therefore, its overall contribution to poverty reduction.

See Also

- ▶ Economic Growth
- ▶ Extreme Poverty
- ▶ Growth and Inequality (Macro Perspectives)
- ▶ Infrastructure and Growth
- ▶ Public Infrastructure

Acknowledgments We thank Jon Temple for useful comments and suggestions. The views expressed here are ours only and do not necessarily reflect those of the World Bank, its Executive Directors, or the countries they represent.

Bibliography

Aghion, P., and P. Bolton. 1992. Distribution and growth in models of imperfect capital markets. *European Economic Review* 36: 603–611.

Aghion, P., and P. Bolton. 1997. A theory of trickle-down growth and development. *Review of Economic Studies* 64(2): 151–172.

Alesina, A., R. Baqir, and W. Easterly. 1999. Public goods and ethnic divisions. *Quarterly Journal of Economics* 114(4): 1243–1284.

Andrés, L. A., B. Briceño, C. Chase, and J. A. Echenique. 2014. *Sanitation and externalities: Evidence from early childhood health in rural India*. *World bank policy research working paper* 6737.

Banerjee, A.V., and A.F. Newman. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101(2): 274–298.

Benitez, D., O. Chisari, and A. Estache. 2003. Can the gains from Argentina's utilities reform offset credit shocks? In *Utility privatization and regulation: A fair deal for consumers?* ed. C. Ugaz and C. Waddams Price. Northampton: Edward Elgar.

Brenneman, A., and M. Kerf. 2002. *Infrastructure and poverty linkages: A literature review (manuscript)*. Washington, DC: World Bank.

Calderón, C., and A. Chong. 2004. Volume and quality of infrastructure and the distribution of income: An empirical investigation. *Review of Income and Wealth* 50: 87–105.

Calderón, C., and L. Servén. 2004. *The effects of infrastructure development on growth and income distribution*. *World bank policy research working paper* 3400.

Calderón, C., and L. Servén. 2010a. Infrastructure and economic development in Sub-Saharan Africa. *Journal of African Economies* 19(S1): 13–87.

Calderón, C., and L. Servén. 2010b. Infrastructure in Latin America. In *The oxford handbook of latin American economies*, ed. J. Ocampo and J. Ros. Oxford: Oxford University Press.

Capuno, J., C.A.R. Tan, and V.M. Fabella. 2011. Do piped water and flush toilets prevent child diarrhea in rural Philippines? *Asia-Pacific Journal of Public Health* 27: NP2122–NP2132. doi:10.1177/1010539511430996.

Chatterjee, S., and S.J. Turnovsky. 2012. Infrastructure and inequality. *European Economic Review* 56: 1730–1745.

Chisari, O., A. Estache, and C. Romero. 1999. Winners and losers from privatization and regulation of utilities: Lessons from a general equilibrium model of Argentina. *World Bank Economic Review* 13(2): 357–378.

Devoto, F., E. Duflo, P. Dupas, W. Parienté, and V. Pons. 2011. *Happiness on tap: Piped water adoption in urban Morocco (manuscript)*. Cambridge, MA: MIT.

Dinkelman, T. 2011. The effects of rural electrification on employment: New evidence from South Africa. *American Economic Review* 101: 3078–3108.

Duflo, E., and R. Pande. 2007. Dams. *Quarterly Journal of Economics* 122(2): 601–646.

Escobal, J., and C. Ponce. 2008. Enhancing income opportunities for the rural poor: The benefits of rural roads. In *Economic reform in developing countries: Reach, range, reason*, ed. J.M. Fanelli and L. Squire, 307–336. New Delhi: Edward Elgar.

Estache, A., A. Gomez-Lobo, and D. Leipziger. 2000. *Utility privatization and the needs of the poor in Latin America: Have we learned enough to get it right? World bank policy research working paper* 2407.

Estache, A., V. Foster, and Q. Wodon. 2002. *Accounting for poverty in infrastructure reform: Learning from Latin America's experience*, *WBI development studies*. Washington, DC: World Bank.

Fan, S., and X. Zhang. 2004. Infrastructure and regional economic development in rural China. *China Economic Review* 15: 203–214.

Ferreira, F. H. G. 1995. Roads to equality: wealth distribution dynamics with publicprivate capital complementarity. *LSE-STICERD Discussion Paper* TE/95/286 (London).

Foster, V., J.P. Tre, and Q. Wodon. 2001. *Energy prices, energy efficiency, and fuel poverty (manuscript)*. Washington, DC: World Bank.

Galiani, S., P. Gertler, and E. Schargrodsky. 2005. Water for life: the impact of the privatization of water services on child mortality. *Journal of Political Economy* 113(1): 83–120.

Galor, O., and J. Zeira. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60: 35–52.

Getachew, Y.Y. 2010. Public capital and distributional dynamics in a two-sector growth model. *Journal of Macroeconomics* 32: 606–616.

Hulten, C. 1996. *Infrastructure capital and economic growth: How well you use it may be more important than how much you have. NBER working paper* 5847.

Jalan, J., and M. Ravallion. 2003. Does piped water reduce diarrhea for children in rural India? *Journal of Econometrics* 112: 153–173.

Khandker, S.R., and G.B. Koolwal. 2007. *Are pro-growth policies pro-poor? Evidence from Bangladesh (manuscript)*. Washington, DC: World Bank.

Khandker, S.R., and G.B. Koolwal. 2010. How infrastructure and financial institutions affect rural income and poverty: Evidence from Bangladesh. *Journal of Development Studies* 46(6): 1109–1137.

Leipziger, D., M. Fay, and T. Yepes. 2003. *The importance of infrastructure in meeting MDGs (manuscript)*. Washington, DC: World Bank.

Lokshin, M., and R. Yemtsov. 2005. Has rural infrastructure rehabilitation in Georgia helped the poor? *World Bank Economic Review* 19(2): 311–333.

López, H. 2004. *Macroeconomics and inequality. World bank research workshop, macroeconomic challenges in low income countries*.

Majumder, R. 2012. *Removing poverty and inequality in India: The role of infrastructure*. *MPRA paper* no. 40941, Department of Economics, University of Burdwan.

Mu, R., and D. van de Walle. 2007. *Rural roads and local market development in Vietnam*. *World bank policy research working paper* 4340.

Navajas, F. H. 2000. El impacto distributivo de los cambios en los precios relativos en la Argentina entre 1988–1998 y los efectos de las privatizaciones y la desregulación económica. In: *La Distribución del Ingreso en Argentina, ed.* Fundación de Investigaciones Económicas Latinoamericanas. Buenos Aires.

Pi, J., and Y. Zhou. 2012. Public infrastructure provision and skilled–unskilled wage inequality in developing countries. *Labour Economics* 19(6): 881–887.

Piketty, T. 1997. The dynamics of the wealth distribution and the interest rate with credit-rationing. *Review of Economic Studies* 64(2): 173–189.

Reardon, T., K. Stamoulis, and P. Pingali. 2007. Rural non-farm employment in developing countries in an era of globalization. *Agricultural Economics* 37(s1): 173–183.

Seneviratne, D., and Y. Sun. 2013. *Infrastructure and income distribution in ASEAN 5: What are the links? IMF working paper* 13/41.

World Bank. 2003. *Inequality in Latin America and the Caribbean: Breaking with history? World bank Latin America and the Caribbean studies*. Washington, DC: World Bank.

Zhang, X., and S. Fan. 2004. Public investment and regional inequality in rural China. *Agricultural Economics* 30: 89–100.

Zhu, N., and X. Luo. 2006. *Nonfarm activity and rural income inequality: A case study of two provinces in China*, *World bank research working paper no.* 3811. Washington, DC: World Bank.

Ingram, John Kells (1823–1907)

R. D. Collison Black

Keywords

Comte, A.; Deductive method; Historical method; Ingram, J. K.; Positivism

JEL Classifications

B31

Ingram's whole professional career was spent at Trinity College, Dublin, of which he became a Fellow in 1846. He subsequently held a remarkable variety of offices there – Professor of Oratory

(1852) and English Literature (1855), Regius Professor of Greek (1866), Librarian (1879) and Vice-Provost (1898) – but was never a professional teacher of political economy.

Nevertheless, Ingram played a notable part in the debates of the 1870s on the future of political economy and became one of the leading advocates in English of the use of the historical method in that science. Ingram's views were initially stated in his presidential address to Section F of the British Association in 1878. Here he attacked the 'vicious abstraction' and attachment to the deductive method of the classical economists, blaming this for the low repute into which political economy had fallen. He advocated the replacement of the deductive by the historical method and that 'the study of the economic phenomena of societyn... be systematically combined with that of other aspects of social existence'. In adopting this approach Ingram was influenced partly by his contemporary T.E. Cliffe Leslie (1826–1882) but chiefly by the positivist philosophy of Auguste Comte, of whom he was an active and lifelong disciple. Of his later economic writings, the best known was, and still is, his *History of Political Economy,* which was for a long time the fullest account in English of the work of the historical school in Germany, France and Belgium. All Ingram's economic work displayed the holistic and normative outlook which he derived from Comte, but did not go far towards fulfilling the programme of historical and comparative studies to which his earlier critique of classical economics pointed.

Selected Works

1878. The present position and prospects of Political Economy. *Report of the British Association for the Advancement of Science,* 641–658. Reprinted in *Essays in economic method,* ed. R.L. Smyth. London: Duckworth, 1962.

1880. *Work and the workman. Address to the Trade Union Congress, Dublin, September.* Dublin: E. Ponsonby.

1888. *A history of political economy.* Edinburgh: A. & C. Black.

1895. *A history of slavery and serfdom.* London: A. & C. Black.

1901. *Human nature and morals according to Auguste Comte.* London: A. & C. Black.

Inheritance

Jack Goody

Inheritance, in the strict sense, is the transmission of relatively exclusive rights at death. Such transmission is part of the wider process of the *devolution* of rights between or within the generations (eventually always between), and particularly between persons regarded as holders and heirs. Devolution continues throughout an individual's life, involving him both as giver and as receiver, and entailing transfers *inter vivos*, between the living for education, marriage, housepurchase, etc. as well as the residuum at death. The connection between inheritance and earlier transfers is given explicit recognition in some customary systems of endowment of sons and daughters where what has already been received is deducted from the final share of the parental estate (as in the revision clause of the Paris–Orleans region from the 16th century). In the same way the trend in European and American tax laws, epitomized in the British Capital Transfer Tax, is to treat as a whole the transfers of property from an estate (in the case of an estate tax) or to one individual or donee (in the case of an inheritance tax).

In a society where production is based upon the household and where rights (whether of ownership, tenancy or use) are vested in the domestic group, then the central importance of the devolution of such rights is clear. This is the case in most sectors of pre-industrial economies, but especially in agriculture and crafts. Where individuals have no such rights in the basic means of production, being employed as wage-labourers or as salaried employees, then the productive system is involved in interpersonal transfers only through

share ownership, the transmission of managerial functions having been 'bureaucratized'; the handing over of such functions takes place at retirement rather than death and involves succession to 'office' rather than inheritance to property.

Thus in industrial societies of whatever political complexion, inheritance is of less significance for individuals and for society (except as windfall income in the first instance and windfall revenue in the second) than in earlier times when, except for the landless, it involved the transfer of rights in the means of livelihood. Even in industrial societies, the state may make special provision for family farms or firms to ensure continuity of the working group.

A radical instance of such a law was enacted in Germany by the National Socialists in 1933, providing for undivided inheritance and forbidding partition by will, the sale of the land or its encumbrance with a mortgage. The law was repealed after World War II, but in Germany as in France and other European countries, the transmission of farm property within the family is protected with the primary aim of ensuring continuity and providing an incentive to work and improve the farm for the next generation.

The Argument About the Inheritance of Wealth

Two divergent views on inheritance are current. On an ideological level, 'socialist' societies, parties and individuals regard inheritance as a way of transmitting inequalities and are therefore in favour of its restriction by taxation or even expropriation. Those espousing 'capitalist' theories look upon the right to transmit acquired wealth to one's offspring as part of the incentive necessary for accumulation, saving and investment. The extreme 'socialist' position is not simply a matter of recent theories of society. At the end of the 18th century the Abbé Raynal declared that at an individual's death, any land he possessed should become a free good. The theme has played a subdominant role in Christian thought over a long period. In 5th-century Gaul, the priest Salvian maintained that since all property came from God, at death it should be returned to his representatives on earth, the Church, for distribution to the poor as well as for its own purposes. Such assumptions left no room for inheritance, so the argument for the social uses of wealth depends not only on the negative case for reducing inequalities but on the positive one for assisting charities. Both positions involve an 'individualistic' view of property, 'freedom' to testate on the one hand and the reduction of the share or relatives (especially of collateral kin) on the other.

At an implicit level, we find a similar spread of ideas in simpler societies. In Africa a distinction is often made between self-acquired property, over which an individual may have a measure of freedom of disposal, and inherited property, especially land, which has come down from his forefathers. In the second case alienation is impossible because an individual is only a temporary custodian, having an obligation (as in some earlier European Laws) to hand the property down in the same line from whence it came. If it had been inherited in the agnatic line, then only agnatic relatives could benefit. In other words the property was 'corporate', or at least 'ancestral'. This notion of an heirloom runs quite contrary to the idea firstly that an individual's wealth should be confiscated in the wider interest either of the government or one of the 'great organizations'; and secondly that he should have completely free disposition over all he has accumulated. Clearly the case for inheritance is more tenable in traditional societies like those in Africa where differences in wealth were small, so that the case for redistribution (motivated either by a positive notion of distributive 'justice' or a negative resentment of inequality) was hardly relevant and where the 'poor' were the responsibility of their kin group. It becomes less tenable with the greater differentiation of capital and income, especially where individuals no longer own the means of production because they are working either for an industrial corporation or for a socialized enterprise, and where it is unnecessary, and often thought undesirable from the bureaucratic standpoint, to attach the next generation to the parental enterprise.

The stress on either the 'socialist' or 'capitalist' pole obviously has some relation to the nature of the ideology and to the organization of productive

enterprise. But it is also true that domestic accumulation and extra-domestic redistribution are aspects of all contemporary social systems. In 1926 the Soviet Union reversed its early position which limited inheritance to small amounts passed on to close relatives or to the surviving spouse, providing there were in need. Later property could be left to anyone and its inheritance, listed as one of the rights of citizens in the constitution of 1936, was seen as a useful incentive to productivity. On the other hand every major 'capitalist' country levies some kind of tax at death, the proceeds of which are destined for the public purse rather than for private enjoyment.

Theories of income distribution often start by assuming normal distribution. This assumption is not adequate for all groups because of the inheritance of property at the death of the parents or other kin. But the main reason lies in the differential interest and capacity of parents to 'invest in' and encourage the abilities of their children. Such encouragement is a kind of transfer *inter vivos*, though it is part of the very process of socialization itself, one in which material gifts may play a major part in helping to provide both shelter in the form of house or apartment and more especially training or capital to generate income.

The ability to transfer privilege from one generation to another is, in the end, intrinsic to family life and to the reproduction process itself with its particularistic interests. Many utopian communities and ideologically based communes attempt to equalize opportunity through early 'schooling' or joint upbringing as well as wealth sharing; in the extreme case parenthood has only a physiological function, upbringing being left to the group. In the extreme case the contradictions become apparent in the longer-term development of communities like the Israeli *kibbutz* where family ties, and hence intra-familial differences, begin to manifest themselves, in limited but significant ways, after the initial period of open recruitment.

Taxation

Whatever their ideological position, all modern societies place a progressive tax on inherited property, a tax that socialist countries see as a means of equalizing advantage, just as attempts are made to counter other benefits derived from family through a national system of education. However, many earlier and some modern taxes were primarily visualized as methods of raising income for the ruler rather than equalizing income among the citizens.

From either standpoint death provides the best moment to raise money, since future beneficiaries are unlikely to offer many objections if they have not yet taken possession. Moreover the property of the deceased often had to be listed, especially under the notarial systems of Europe, as part of the process of handing over: consequently the basis for an assessment already existed. Such forms of taxation have a long history, appearing in Rome as the *vicessima hereditatum*, the twentieth penny of inheritance. In feudal Europe the heriot, payable at the transfer of an estate, accrued to the lord; but already in 1694 a central death tax was introduced in England, taking its modern form in 1779–80. Taxes on inherited wealth thus long preceded taxes on income.

Ways of avoiding tax also had a long history. Under English law, trusts and life-estates could be set up in order to skip a generation in the transfer of property. Discretionary trusts were not available in continental Europe. But other avoidance measures included the handing over of property *inter vivos*, a long-established tradition (even extending to the farm itself), although such a practice now runs up against taxes on gifts or capital transfers: a modern alternative, that of changing one's country of residence, is more difficult to control, except by controlling the total outflow of capital from the country.

Today these possibilities remain relatively little used, and yet the revenue role of death taxes is not great. No taxes, remarks Shoup (1968, p. 559), have had a better reputation to less effect. This is partly because of avoidance and high rates of exemption, but mainly because individuals divest themselves of property to their children or use it for support in their old age, which is especially easy when personal property relates to consumption rather than to production. However, in the USA where charitable gifts are exempt, private foundations reap important benefits, while in the UK

major contributions to national collections of art, buildings or land result directly from such taxes.

The History of Inheritance

The system of inheritance, involving the transmission of relatively exclusive rights over material objects, clearly varies with the mode of production. In hunting and gathering societies, rights of this kind are minimal; much of a man's property may be destroyed at his death, each individual fashioning himself or acquiring from others the tools he requires for his own use. The destruction is an aspect of the close identification of a person with the property he has created or used that is characteristic of such societies.

In simple agricultural economies, rights in land become elaborated, although with shifting cultivation access is more important than ownership. Access to the basic means of production is likely to be achieved through membership of a kin group (by descent or affiliation) rather than through an inheritance transaction. But other types of property, livestock, houses and exchange items, are transmitted in the course of long funeral ceremonies.

Where animal or other forms of non-human energy can be harnessed in the process of production, land becomes a scarcer resource, more differentiated in its distribution, with a greater complexity of rights, 'ownership' tending to be the prerogative of the dominant groups, and tenancy (or even labouring) the prerogative of others. In the case of tenancy it is landlords that tend to make the rules for the transfer of property, insisting for example on indivision (keeping the holdings intact) or on redistribution (keeping the holdings equal). The system of inheritance itself is influenced by the existence of stratified access to land, each group attempting to employ strategies of heirship to maintain or improve their position. The situation with regard to stratified access to livestock for pastoral peoples is different in certain important respects (the herd is more easily divided, increased and consumed) but tends to produce broadly similar strategies as are found in plough agriculture.

In industrial societies the situation is radically different because the vast majority of the population labour for wages rather than owning rights in the means of production. As we have seen, inheritance consequently plays a very different and more peripheral role.

Everywhere inheritance is basically a kinship transaction. While other persons may be involved, the core relationships are close 'familial' ones. In the simpler societies eligible kin are rarely, if ever, lacking since virtually all relationships are between kin. In complex ones, the definition of eligibility tends to be narrower, friendship supplements kinship, the percentage of unmarried tends to be higher and in any case other institutions, the 'great organizations' of church, state, as well as the charitable foundations, make their own demands; nevertheless the 'family' continues to dominate the process of the transfer of wealth between generations.

In kinship terms one can transmit laterally to spouses or siblings, or lineally to children or to siblings' children: the choice is one of priority since all ultimately has to go to the next generation. Downwards transmission for men can be to the sister's child (uterine inheritance) or to own children (agnatic inheritance). In simple hoe agricultural societies inheritance between spouses is rare, transmission tending to be homoparental, that is, male to male, female to female. Such is the case in much of Africa where economic differentiation is relatively small and access to land available to all or most free individuals. Historically, these forms of inheritance were usually associated with the presence of unilineal descent groups (clans or lineages) in which property is transmitted between its members of the same sex.

Patrilineal and matrilineal clans with agnatic and uterine systems of inheritance are found in all types of pre-industrial society but matrilineality is more frequent with tropical hoe agriculture (in which women often do much of the farming, continuing their role as food gatherers in hunting societies) while patrilineality predominates when agriculture is combined with the herding of large livestock (whether or not these are used for plough traction).

The alternative form of inheritance, dominant in one form or other since the advent of plough agriculture, is diverging, or bisexual, that is, with children inheriting from both parents, and parents transferring wealth to daughters as well as to sons, but not necessarily at death. One form of early transfer is the direct dowry whereby daughters are 'endowed' when they depart at marriage. While the man-to-man (homoparental) transmission of Africa excludes inheritance by spouses, diverging devolution in Eurasia tends to exclude uterine inheritance by sister's children, concentrating on passing property, after the surviving spouse has been taken care of, directly to one's own 'natural' children, and even encouraging the adoption of outside heirs before allowing property to go to collaterals. The elementary family takes precedence.

Differences in stratification associated with advanced agriculture work in favour of the identification of conjugal statuses. Transmission in such societies is usually bisexual. At marriage some kind of conjugal fund (or identity of interest) is established and the property is transmitted, though not in equal proportions, to the children of both sexes, with certain types tending to be sex-linked; for example, land may be passed down to males alone where it is associated with male status among the nobility, as under the law of the Salian Franks. Since handing down occurs not only at death but on earlier occasions, especially at marriage, questions concerning the equality of 'inheritance' have to be looked at in terms of the total process of devolving property between holder and heir throughout their lifetime. For example, a woman may receive less at death because she has received a larger dowry at the time of her marriage, even as the promise of dower to maintain her as a widow.

The different treatment of siblings depending on birth order takes the form of primogeniture, ultimogeniture, or partition, known to earlier English law as Borough French, Borough English and gavelkind respectively. Rarely if ever does one find the transmission of the entire conjugal estate to a single sibling but rather the preferential treatment of one at the expense of the others. Such a preference may be tied to particular obligations, as when the inheriting child is expected to stay with the parents in their old age. In other cases the preference for one child (unigeniture) is related to the desire to keep the family estate intact, either because it will only support one family (among the poor) or because it is tied in with status consideration (among the rich). The first situation applied to pre-Revolutionary China where the poor tended to live in stem households (containing one member of each generation) while the richer lived in larger, extended ones. The poor either had less children *in toto* or the additional offspring migrated elsewhere or worked locally as landless labourers. The second situation was found in parts of 'feudal' Europe where title and position were linked to estate and income; just as one child succeeded to the title or office, so he had to inherit the bulk of the estate to which it was attached. Younger sons sought their fortunes elsewhere in the great organizations of the church or the army, to which they had access as a consequence of the political power assured by the parental estate.

Devolution, Retirement and Interpersonal Conflict

Until the end of the 19th century (and still today in some areas of rural Europe), propertied classes endowed their daughters at marriage (and sometimes on entering a convent, on becoming 'a bride of Christ') with part of the 'portion' they would otherwise have inherited at the death of the parents. In some farming communities the parents would hand over the farm to their son or daughter on the occasion of their marriage, reserving for themselves certain rights to bed and board which were sometimes embodied in specific retirement contracts. One of the penalties of such an early handing over of property is that parents are placed in a King Lear situation, overly dependent upon the succeeding generation and running the danger of neglect (or 'ingratitude'). On the other hand to hang on till the end to property that is critical to status or survival leads to the opposite kind of tension characterizing the Prince Hal situation, where the son attempted to grasp his father's

crown while he was still alive. These problems are of less significance in wage-earning societies where individuals are more dependent on income than on capital, and it is into training children for future employment that parents invest their time and wealth, rather than devolving property at marriage or even at death. While the state provides a minimum level of support, wealth may be invested in a pension or retained by the elderly for their support, possibly disappearing with their death in the form of annuity. Little conflict arises between holder and heir, who rarely continue to reside together (except in the case of spouses); inheritance tends to come late and to be seen as a 'windfall'. Its distribution may still give rise to conflicts within the group of potential beneficiaries, while even the prospect of a windfall produces enough underlying tension to fill the pages of many a piece of detective fiction (and 19th-century classics like *Middlemarch*), although significantly the plots are frequently located in the past when greater weight attached to rentier income.

Rights Transferred

The rights transferred from the dead to the living are largely those in material property, houses, land, money, heirlooms, but they may also include rights to receive rent, interest, dividends from shares. In earlier societies they included rights to the services of other humans (of serfs and slaves), even to women as wives and to men as husbands, as in the Jewish practice of leviratic inheritance, taking on the widow of a childless brother with a view to breeding offspring to his name. Any semblance of the inheritance of widows was rigidly excluded from the law of England at the time of the Reformation since it was on grounds of the invalidity of such marriages that Henry VIII set aside his first wife (Catherine of Aragon, the widow of his dead brother, Arthur) in his search for a successor and an heir. Inheritance may also involve other types of right, those of a non-corporeal kind, right to songs and stories (copyright), rights to armorial bearings, titles, etc., although here we touch upon the field of succession to social position, to office, to nobility or to similar benefits.

Inheritance is not only concerned with rights; duties too are involved; debts have to be paid from the estate; the acceptance of an inheritance may involve a change of name, of residence and even, especially in societies that resolve disputes by means of feud, of the specific obligation to settle a score.

A broad distinction is made in Anglo-American law between real and personal property, roughly between land a chattels, between movables and immovables. A similar distinction is found in many other cultures and is related to the special position of land within the general category of property, since it acts both as a factor in production and as a locus for all social activity. Hence different rules and practices are applied to these two categories; in England after the Norman conquest it was the royal courts that dealt with real, and ecclesiastical courts with personal property, the former emphasizing indivision, the latter allowing more testamentary freedom and alienation. Land was subject to different rules until 1926 and its transfer is still hedged about with formalities that mark it off from all other forms of property. For a hierarchy of rights is always involved; these may refer to usufruct, tenancy, mortgage, metayage, and a host of arrangements (including sovereignty itself) to which other property is not subject and some of which need to be acknowledged in the deed of transfer itself.

Testate and Intestate

There are two types of inheritance in literate cultures, testate and intestate. The former involves making (writing) a will or testament, the latter describes what happens when there is no such statement. In nonliterate societies inheritance is automatically intestate, and 'custom' lays down how property should be distributed. There is little 'freedom' to alienate goods from the recognized heirs so that even gifts *inter vivos* have to be monitored.

The written will or testament introduces a measure of certainty in situations, reducing possible

conflict or indeterminacy; where this is available, nuncupative (that is, oral) wills are considered valid only under exceptional circumstances. But in early times one of the main functions of the testament was to certify that any alienation from customary heirs was according to the wishes of the deceased. In other words its very existence assumed a degree of 'freedom', by which is meant freedom of choice for the holder, limiting the right of 'society' to say who should be the recipient. It is not surprising that the will with its corresponding freedom of testation was encouraged by the early Christian Church as a way of acquiring property to be put to divine purposes. And in more recent times it has been a central instrument for the transfer of wealth to charitable foundations. Without its intervention, inheritance goes to the family.

Testamentary inheritance occurs by means of the written will, although initially the latter term applied only to real property (land), the testament to personal property. Literacy is thus essential either on the part of the testator or on the part of the notary, lawyer or priest who draws up the will. The fact that the document has to be proven in court means that professionals tend to be employed for the purpose. Hence the whole industry of literate legal specialists involved in writing the will, in helping it to come into effect and in administering the estate, the last being currently the most profitable part of the enterprise in anticipation of which other charges may be scaled down. It is these specialists who help to ensure that the formalities are observed, not only in the words but in the witnesses, and that the will is not invalidated for other reasons. All of which tends to take the mechanics of transfer out of the hands of kin, and places the process firmly under the charge of those who engage in it for their livelihood and who tend to create their own specialist language, codes and organization.

Testamentary Freedom

In oral societies, little scope existed for alienation from the heirs who were regarded as the proper recipients. On the other hand testamentary freedom has disinheritance as its corollary. The problem of exclusion became acute in the early days of Christendom since some religious advisors encouraged the old to leave all to the Church and nothing to their kin. The Church itself, and later 'hell fire' and 'charity-begins-at-home' statutes, legislated against such forms of disinheritance and indeed most contemporary systems reserve a 'legitimate' part for the spouse and the children. In this way testamentary freedom is limited by law so that close kin benefit from a portion of the estate, although not to the same extent as under intestacy. This limitation holds even in ancient Roman and modern Anglo-American law where freedom to disinherit was greater than in the civil law regimes of the continent. In England, the obligation to leave a minimum share of personal property to close kin disappeared in the course of the 17th and 18th centuries, while in 1833 the widow lost her right to a dower. In Scotland, on the other hand, rights such as the bairn's part continued. Indeed in wealthier families in England these rights were always maintained by entails, by the strict family settlement of the 18th century and by earlier devices which prevented the splitting of the estate, while parallel practices deriving from late Roman law existed on the continent. Such arrangements were the subject of objections by some because they kept land from the market and made it impossible to raise a mortgage to effect improvements. In Europe the system collapsed with the French Revolution, following which the Napoleonic Code tried to ensure partition. But in England it persisted until the Settled Land Act of 1882. More recently Family Provision Acts have restored some of the protection given to the surviving spouse and to the children, and in the Soviet Union to anyone previously dependent upon the deceased.

In fact, the beneficiaries of inheritance under a will do not turn out to be greatly different from 'intestate' inheritance, not only because of legal restrictions but because the contents of written wills follow the general sentiments of donors. Indeed because of its flexibility the pattern of

testamentary inheritance may be closer to the moral climate of opinion, as in the preference it gives to the 'spouse-all' provisions of modern Anglo-American law, whereas division with the children obtains in the more conservative case of intestacy. The legal formalism connected with literacy 'tends to generalize rules that have originated in connection with special situations into applications beyond their initial scope' (Rheinstein 1974, p. 590). At the same time the written rule tends to preserve past situations so that intestacy laws have 'frequently looked obsolete, confused, or arbitrary'.

Inheritance Under Current Anglo-American Law

Intestate rules in Anglo-American law usually split the property between the surviving spouse and the children. When people make wills, on the other hand, they use the testamentary freedom to leave all to the spouse, usually the wife as she is often younger and lives longer than the husband. In general it is women as widows that benefit most from inheritance. Only after the widow's death does the property drop a generation. The one exception is in the case of a remarriage where specific provision is often made in advance for the children of 'the first bed' in whose welfare the surviving spouse may have less interest.

Even here, despite the potential difference between the outcome of testate and intestate inheritance, the results are very similar since children normally hand over the portion to which they are legally entitled to their parent so that he or she can continue to lead an independent life. When the next generation eventually inherits, the property is usually split equally between children regardless of sex. However, there is one major exception to equality of partition. When one of the siblings has looked after the parents in their old age, testamentary 'freedom' or intestate adjustment is used to allocate that person a preferential share. This was one of the roles of preferential primogeniture or ultimogeniture in early English law, the last-born son being known in some parts as the *astrier*, the one who remains by the hearth. Otherwise equality is the norm both in law and in practice. Whatever discrimination operates against women in other sections of the society, little is now manifest in testamentary matters, either as spouses or as daughters (however, the 'poor' widow who did not produce a dowry can be helped from the estate, both in Justinian's law and in modern Lousiana). A wife tends to regard an inheritance as her personal peculium, a nest-egg. Given the relatively late age that most people receive legacies, these may make little difference to the lifestyles of the recipients, who sometimes use them as gifts *inter vivos* to assist their own children rather than themselves.

See Also

- ▶ Economic Anthropology
- ▶ Family
- ▶ Inheritance Taxes
- ▶ Ricardian Equivalence Theorem

Bibliography

Goody, J. 1962. *Death, property and the ancestors*. Stanford: Stanford University Press.

Goody, J., J. Thirsk, and E.P. Thompson (eds.). 1976. *Family and inheritance: Rural society in Western Europe 1200–1800*. Cambridge: Cambridge University Press.

Renner, K. 1929. *The institutions of private law and their social functions*. Trans. from German, London: Routledge & Kegan Paul, 1949.

Rheinstein, M.Y. 1974. Inheritance. In *Encyclopaedia Britannica*, vol. 19. Chicago: Encylopaedia Britannica.

Shoup, C. 1966. *Federal estate and gift taxes*. Washington, DC: Brookings Institution.

Shoup, C. 1968. Taxation: death and gift taxes. In *International encyclopedia of the social sciences*, vol. 15. New York: Macmillan.

Sussman, M.B., J.N. Cates, and D.T. Smith. 1970. *The family and inheritance*. New York: Russell Sage Foundation.

Wedgwood, J. 1929. *The economics of inheritance*. London: G. Routledge & Sons.

I

Inheritance and Bequests

Kathleen McGarry

Abstract

The importance of bequests, their role in capital accumulation, and the motivation behind these transfers has long been the subject of debate among economists. Various models of intergenerational transfers yield different predictions about the responsiveness of bequests to changes in incomes of the donors and recipients and thus to the impact public policy. Yet, despite the intuitive appeal of these models, none has proved to be consistent with empirical patterns. This article discusses the alternative theories of transfer behaviour, examines the empirical work testing their predictions, and discusses the role of estate and gift taxes in affecting bequest behaviour.

Keywords

Accidental-bequest motive; Altruism; Annuities; Bequest motive; Bequests; Charitable contributions; Consumption smoothing; Estate taxation; Exchange motive; Gift taxation; Health insurance; Inheritance taxation; Inheritances; *Inter vivos* transfers; Intergenerational transfers; Life insurance; Marginal utility of consumption; National Longitudinal Surveys (NLS); Pensions; Savings behaviour; Social Security (USA); Succession laws; Tax avoidance; Transfer taxation; Wills

JEL Classifications

J10

Fascination with inheritances and bequests began long before economists formalized models of transfer behaviour. Literature and history are rife with examples of the role of inheritances (for example, Shakespeare's *King Lear*). Societies have laws governing bequest behaviour and governments have long employed bequest, gift and/or inheritance taxes (jointly termed 'transfer taxes') as a means of raising revenue. Economists, in turn, have examined the motivation behind bequests and their importance in driving economic behaviours. These transfers have been theorized to play a central role in the accumulation of wealth, the degree of inequality present in a society, and the interactions among generations. This article touches briefly upon several economic dimensions of inheritances and bequests.

Although the focus in here is on inheritances, the distinction between bequests, made at the time of death, and *inter vivos* transfers, made during life, is somewhat arbitrary. Intended bequests may be made prior to death as a means of reducing estate or inheritance taxes, avoiding other legal requirements pertaining to the settling of an estate (such as probate), or alleviating liquidity constraints and smoothing the consumption of an intended heir. Conversely, resources transferred during life could well have been saved and transferred at death in the form of a bequest. Indeed, much of the literature attempting to assess the importance of bequests in contributing to the capital stock has included the magnitude of *inter vivos* transfers along with bequests in any calculations. Similarly, economic models of the motivation for bequests are generally applicable to *inter vivos* transfers as well. Finally, in many cases, transfer taxes apply to both *inter vivos* transfers and bequests.

In this discussion I focus primarily on bequests but, where appropriate, I draw on the research examining *inter vivos* transfers as well. I use the generic term 'transfers' to refer to either bequests or *inter vivos* transfers. Also, for ease of exposition I occasionally refer to donors as parents and recipients as children. Obviously, bequests are frequently made to non-child heirs, but the use of this terminology makes the discussion less abstract and also accurately reflects the situation for the majority of bequests.

Much of the literature examining bequests has sought to assess the relative importance of inherited wealth and life-cycle savings as components of the existing wealth stock. Estimates of the

relative importance of bequests have varied widely. Numerous researchers have put the fraction of wealth due to transfers at 15–20 per cent (see Modigliani 1988, for a discussion), but some studies argue that the figure is much higher, concluding that transfers account for a large share of wealth holdings (for example, Kotlikoff and Summers 1981). Although the existing estimates bracket an extremely large range, even the lower figures indicate that these transfers are an important economic phenomenon and crucial to understanding patterns of savings and life-cycle behaviour. Furthermore, inheritances and *inter vivos* transfers can potentially have substantial impacts on the well-being of the recipient, his economic behaviour, and on broader measures of the distribution of income and measures of inequality.

Intentional Versus Accidental Bequests

The importance of bequests and their impact on macroeconomic measures such as saving rates and individual well-being depends to a great extent on the motivation driving the transfer. One school of thought argues that bequests are accidental, the result of an uncertain length of life. Individuals save to finance consumption during their retirement years and whatever wealth remains when they die is bequeathed to their heirs (Davies 1981). Because they do not know how many years of consumption they must finance and do not want to exhaust their resources prior to death, individuals will typically die with some amount of wealth. Hurd (1987) tests the data for consistency with an accidental-bequest motive. He argues that, if bequests were intentional, one would find that individuals who had a strong bequest motive would dissave at a slower rate than those with a weak bequest motive. As a proxy for the strength of a bequest motive, Hurd uses the presence of children. He finds no difference in rates of spend-down of assets for those with and those without children, and thus concludes that there is no operative bequest motive: observed bequests are the result of an uncertain date of death.

The uncertainty in this accidental-bequest scenario need not arise solely from uncertainty about the length of life, but could stem from a variety of sources: an individual might conserve assets to guard against expenses arising from a negative health shock, the need for long-term care, or uncertainty about returns on investment.

Additional corroborating evidence for the notion of accidental bequest comes from the failure of many individuals to specify a particular distribution of their estates. Although laws in the United States, as in much of the world, allow an individual to distribute his estate in any way he wishes through the use of a will, the use of wills to divide resources is far from universal. An individual who dies without a will is said to have died intestate. In these cases the assets of the deceased are distributed according to the laws specific to the area in which she resided. In the United States, laws differ by state but earmark a large fraction of the estate for a surviving spouse, followed by children, grandchildren and parents, with shares equally divided within kinship category. Although the reliance on succession laws to distribute assets suggests that the individual may not have thought about bequests and therefore does not have a bequest motive, it can also be argued that, if the succession laws mirror (or come close to) the distribution that the deceased would have chosen herself, a reasonable person might forgo the trouble and expense of writing a will and allow the state to divide her assets. Because wills, when they do exist, typically divide estates equally among children, as do succession laws, the failure to execute a will may indeed reflect a satisfaction with the default distribution.

The chief criticism of the notion of accidental bequests is that an individual who is concerned about uncertain future expenses could instead purchase insurance protecting against these expenditures. Insurance against outliving one's assets is available in the form of annuities which guarantee a stream of income for life and eliminate the possibility of dying with unspent wealth; instruments such as health insurance and long-term care insurance can protect against other types of unplanned expenditures. The accidental-bequest

motive thus requires that these insurance markets function imperfectly.

Indeed, the potential for annuities to eliminate the possibility of an accidental bequest has been used to test for a bequest motive. If all wealth is annuitized, one has nothing to leave as a bequest. If complete annuitization is not optimal and a bequest is desired, an individual can convert a portion of his annuity income into a bequest by purchasing a life-insurance policy. Thus, life insurance can be used to offset the effects of an annuity. 'Over-annuitization' may not be uncommon; the prevalence of mandatory old age pensions (either public or private) suggests that many workers may retain a substantial portion of their retirement resources in annuities, perhaps more than they would choose. Bernheim (1991) examines the relationship between annuitization, in the form of US Social Security benefits, and the holdings of life insurance and private pensions. He finds that, conditional on lifetime resources, those with a greater Social Security benefits hold more life insurance and somewhat smaller private pensions. This result suggests that these tools are used to de-annuitize wealth and support the notion of an operative bequest motive. Indeed, the extensive life-insurance holdings observed in the population provide prima facie evidence that individuals are sufficiently concerned about the well-being of their heirs that they are willing to reduce own consumption.

Another argument against the accidental-bequest motive comes from the growing literature on *inter vivos* transfers. The large number of *inter vivos* transfers observed in the data are unquestionably intended and suggest that bequests might likewise be intentional.

Motivation for Bequests

If individuals intentionally leave bequests, the next question is: why? What motivates an individual to forgo consumption in order to leave assets to his heirs? Several behavioural models have been offered to address this question, but the results of empirical tests remain inconclusive.

Perhaps the most obvious explanation for the existence of bequests is that donors are altruistic; they care about the well-being of their heirs. The standard specification of the altruism model (Barro 1974; Becker 1974) includes the heir's utility as an argument in the utility function of the donor. Formally, the utility of the donor (say, parent), U_p, is written as

$$U_p = U\left(C_p, U_k(C_k)\right)$$

where C_k is the consumption of the heir (say, child). (Note that this formalization is based on very specific assumptions and thus has implications that may differ from what one might more generally regard as altruistic behaviour in other contexts; Pollak 2003.) Consumption for p and k depend on the resources of each party prior to the bequest and on the size of the bequest. In this specific formulation the donor will make transfers until the marginal utilities are equalized across arguments of the utility function. Because the marginal utility of consumption is assumed to be decreasing, bequests will increase with the income of the donor but decline with increases in the income of the (potential) recipient. If there is more than one child, the parent will endeavour to equalize the marginal utility of consumption across children. Again, because marginal utility is decreasing in consumption, less well-off children will receive larger bequests. Thus, within a family, bequests will be compensatory and will serve to mitigate inequality.

Alternatively, transfers may be part of an exchange regime wherein the donor reimburses the recipient for specific services or behaviours. A parent compensating a child for providing home health care or simply for paying attention to the parent (Bernheim et al. 1985) would be an example of possible exchange-related transfers. In this case the donor's utility function has as its arguments her own consumption and the goods or services 'purchased' from the child. Formally,

$$U_p = U\left(C_p, S_k\right)$$

where S_k is a measure of services provided to the donor. The price of the services depends on the

price of the recipient's time, with services purchased from high-income individuals being more costly than those purchased from low-income individuals. As the price of the good or service increases, the quantity purchased declines. In this case, then, the parent will be less likely to purchases services from a high-income child and the *probability* of a transfer will decline with the income of the child. However, conditional on purchasing services, the relationship between the transfer and the income of the recipient is indeterminate: the total amount of the transfer, price multiplied by quantity, can either rise or fall with the income of the heir, depending on the relevant elasticities.

Several other models have been discussed in the literature but have received less attention. A 'paternalistic' model argues that parents care not just about the utility of their children but about their actual consumption bundles. In this case a parent might bequeath money to a child through a trust specifying that it be used for certain purposes, such as schooling, or available only at certain ages when the parent believes the child's preferences will more closely mirror her own.

A 'warm glow' model posits that donors receive utility from the act of giving itself and not from the impact the gift has on the utility of the recipient (Becker 1974; Behrman et al. 1982; Andreoni 1989). Such a model might be relevant in the decision to make charitable gifts, wherein the donor is unlikely to observe the increase in utility accruing to the beneficiary as a result of the donation, yet she derives satisfaction from making the gift.

A good deal of research has attempted to discern which of the models best represents observed behaviour. The models are typically written in a static one-period framework and in such a case testing the altruism model is straightforward. Simple tests of the relationship between the probability and amount of the transfer on the one hand and the income of the potential recipient on the other should reveal a negative relationship: that is, transfers should be compensatory. However, there is a stricter test of the altruism model based on the magnitude of the response to variations in the incomes of the donor and the recipient. Specifically, the model requires that, conditional on transfers being made, an increase of one dollar in the income of the donor, accompanied by a decrease of one dollar in the income of the recipient, must be met by an increase of one dollar in the amount of the transfer (Cox 1987). This test imposes a strict 'adding up' constraint on the estimated coefficients on the donor's and the recipient's income variables in a regression equation for the amount of a transfer (conditional on a positive amount). In contrast to this strict test of the altruism model, nearly any relationship between income and transfers is possible in an exchange regime. This ambiguity makes it difficult to discredit the exchange model. Not only can the relationship between the income of the recipient and the amount of the transfer go in either direction, but the components of the exchange need not be made coincidently, making it difficult to observe both sides of the transaction in data.

Observed Patterns

Although *inter vivos* transfers and bequests appear to be substitutes to some extent, the two forms of giving exhibit strikingly different patterns. *Inter vivos* transfers have nearly uniformly been found to be compensatory, with more going to the less well-off children. This negative relationship between the income of the recipient and the probability and amount of a transfer is consistent with the altruism model, but is also consistent with an exchange regime wherein the donor purchases more services from lower-income heirs. Where the strict test for altruism based on the relationship of the income derivatives (that is, the magnitude of the responsiveness of transfers to changes in the incomes of the donor and recipient) has been applied, however, it has failed decisively, with estimated responsiveness closer to zero than to the value of 1 predicted by the model (Altonji et al. 1997).

Perhaps the seminal article testing for the existence of an exchange motive is Bernheim et al. (1985). In that paper the authors hypothesize that parents hold bequeathable wealth and use the

possibility of disinheritance to elicit desired behaviour from their children. The study finds a positive correlation between parental bequeathable wealth and the amount of attention children pay to their parents. Recent work has questioned the empirical results (Perozek 1998) but the notion of a parent reimbursing a child for the provision of care or other behaviour has some appeal as does the idea of an altruistic parent using bequests to compensate a less well-off child.

Although economists often shy away from directly questioning individuals about their motives, one method of attempting to discern the motivation behind the division of bequests is to ask parents about their intentions. The National Longitudinal Surveys (NLS) included such questions, explicitly asking those respondents who reported that their wills provided for unequal division of their estates why they were allocating their assets in such a way. Light and McGarry (2003) examine this question and find that motives based on altruistic concerns and those based on some sort of exchange were of nearly equal importance.

Despite the predictions of the altruism and exchange models and the compensatory transfers observed for *inter vivos* giving, examinations of both actual bequests and existing wills find that equal division among children is the norm. Some of the first work in this arena found evidence that bequests were compensatory (Tomes 1981), but other work appeared to contradict this conclusion – for example, Menchik, (1980). More recent studies have found overwhelming evidence that estates are typically equally divided. Wilhelm (1996) uses a sample of US estate tax returns and finds that two-thirds of decedents with two or more children divided their estate exactly equally among the children and three-quarters used a division in which inheritances differed by no more than two per cent from the within-family average. Although Wilhelm's study is necessarily limited to decedents whose estates filed a tax return and who were therefore in the upper tail of the wealth distribution, similar results have been found for the general population. McGarry (1999) examines reports about existing wills for those who are still living and finds that more than 80 per cent of respondents report that their will divides their estate 'approximately equally' among their children.

This equal division is difficult to reconcile with either the altruism or the exchange model, both of which predict a correlation between the income of the recipient and the magnitude of the bequest. This empirical regularity has thus led several authors to propose alternative models of behaviour. Wilhelm (1996) and Bernheim and Severinov (2003) posit that unequal division is costly to parents in that they foresee that such a division could lead to unhappiness on the part of the children/intended heirs. If the difference between the utility obtained through an equal division and that obtained through an unequal allocation is greater than the utility cost (in terms of unhappy heirs) of unequal division, the parent will simply divide her estate equally among her children. McGarry (1999) provides an alternative model wherein the parent's uncertainty about the future incomes of her children lead her to resort to equal division, except in cases where large differences in the future incomes of children are expected.

Transfer Taxes

Bequest and inheritance taxes have an extremely long history, dating back thousands of years, and can arouse strong feelings. Historically these taxes have been imposed as a revenue-raising mechanism, often in times of war and as a means of diluting the concentration of wealth (see Johnson and Eller 2001, for a discussion of the history of estate taxes). In the United States the modern estate tax was implemented in 1916 to help finance the war effort (Joulfaian 1998). Although the fraction of estates owing a tax has varied over time, it has typically been small, hovering around two per cent. The future of the estate tax in the United States is uncertain. Under current law the tax is being gradually phased out, to be completely eliminated in 2010 but reinstated in 2011.

The form transfer taxes take varies across countries. In the United States taxes are levied on bequests and gifts, with the tax rate applied, broadly speaking, to the total value of the transfer regardless of how it is divided, although various aspects of the tax code lead to important differences in the cost of the two types of transfers (Jolfaian 1998). Transfers to spouses and charitable organizations are exempt from tax. Not all governments have used the same approach as that employed in the United States. Many countries instead have enacted inheritance taxes wherein the tax owed depends on the amount received by an individual heir and often on the legal relationship between the decedent and the heir. These different tax bases, and the particular rules governing the evaluation of the transfers, produce varying incentives for the distribution of estates and gifts. However, the specific behavioural responses also depend on the motivation behind transfers. Although uncertainty exists about this motivation and thus about some of the predicted effects, numerous studies have shown that transfers (both *inter vivos* transfers and bequests) are responsive to tax rates (for example, Bernheim et al. 2004; Joulfaian 2005), and there exist sizable segments of the financial and legal industries devoted to estate planning (that is, reducing estate and gift tax liabilities; see Cooper 1979, for a fascinating look into methods for tax avoidance).

Despite these findings, and the public sentiment against the tax, several empirical studies have shown that some of the simplest tax avoidance schemes often go unexploited. For instance, in the United States *inter vivos* gifts of less than a given amount in any specific year are exempt from gift tax and can thus be used to 'spend down' a potentially taxable estate. Despite this opportunity, at least half of those whose estates appear likely to incur estate tax do not make such transfers (Poterba 1998). Numerous hypotheses have been proposed to explain the failure to make 'early bequests', including the fear that the resources will be needed at some future date, the utility obtained from holding wealth, or the mistrust of children and their ability to manage the funds. None of these explanations appears sufficient to explain this behaviour fully.

Charitable Giving

Bequests are made not just to individuals but often to charitable institutions. In the United States the tax-exempt status granted to charitable bequests reduces the price of donations to these sorts of organization relative to the price of giving to other non-spousal heirs. Numerous studies have found that the lower tax price substantially increases charitable donations. A recent study by the United States Congressional Budget Office estimates that total charitable bequests would decline by 6 to 12 per cent in the absence of the estate tax, an amount similar to the range of estimates produced by various studies over the years (U.S. Congressional Budget Office 2004).

Behaviour of Heirs

Much of the research assessing the importance of bequests in affecting economic behaviours has focused on the behaviour of the donor, the motivation for the transfer, the response to estate and gift taxes, and the effect of desired bequests on savings behaviour. Less frequently examined is the economic response of the heirs.

From the point of view of the heir, inheritances increase financial resources and would therefore be expected to increase the consumption of normal goods, including leisure. The potential reduction in the labour supply of heirs is often cited as a motivation for a transfer tax (for example, Carnegie 1962). Despite the theoretical implications, the several studies examining this issue have found relatively small negative effects on earnings of workers (for example, Joulfaian and Wilhelm 1994). The small responses are not surprising in that the distribution of bequests is extremely skewed, so that for most heirs the amounts received are small relative to their lifetime incomes. Furthermore, as recent work in labour economics has

demonstrated, it may be difficult to adjust hours of work on the margin. Indeed, evidence of a negative labour market effect is somewhat larger with respect to whether one participates in the labour force at all (Holtz-Eakin et al. 1993), suggesting that the length of the working life may be a dimension along which adjustments are more easily made.

Finally, if bequests are fully anticipated, their effect on desired hours of work ought to have been already incorporated into behaviour, and there should be no discernible response at the time the heir receives the inheritance. Thus, only unanticipated bequests or bequests received by previously liquidity-constrained heirs would be expected to spark a change in behaviour. As evidence of the potential importance of liquidity constraints, Holtz-Eakin et al. (1993) find that inheritances can spur entrepreneurial activity.

Conclusion

Bequests play a central role in numerous economic models and as such have long attracted the attention of economists. The strength of the desire to leave bequests and the motivation behind these transfers have direct implications for such fundamental behaviours as life-cycle savings and consumption. From a public policy point of view, bequests affect the accumulation of the capital stock, the distribution of income, and the ties across generations. They also provide a source of tax revenue. Furthermore, estimates suggest that an enormous amount of wealth could be bequeathed in the coming decades, making the issues quite timely.

In attempting to understand the motivation behind bequests economists have offered several theoretical models, all of which have some intuitive appeal. Although no agreement has been reached on the most plausible theory or their relative importance, the recent availability of richer data-sets and the use of administrative records provide some hope that patterns of intergenerational transfers will be better understood. Gaining insight into the motivation behind transfer behaviour will help us to assess the potential impacts of tax policies and public transfer programmes, and to understand more completely the impact of population ageing.

See Also

- ▶ Altruism, History of the Concept
- ▶ Estate and Inheritance Taxes
- ▶ Intergenerational Income Mobility

Bibliography

Altonji, J., F. Hayashi, and L. Kotlikoff. 1997. Parental altruism and inter vivos transfers: Theory and evidence. *Journal of Political Economy* 105: 1121–1166.

Andreoni, J. 1989. Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* 97: 1447–1458.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1063–1094.

Becker, G. 1974. A theory of social interactions. *Journal of Political Economy* 82: 1063–1094.

Behrman, J., R. Pollak, and P. Taubman. 1982. Parental preferences and provision for progeny. *Journal of Political Economy* 90: 52–73.

Bernheim, B. 1991. How strong are bequest motives? Evidence based on estimates of the demand for life insurance and annuities. *Journal of Political Economy* 99: 899–927.

Bernheim, B., R. Lemke, and J. Scholz. 2004. Do estate and gift taxes affect the timing of private transfers? *Journal of Public Economics* 88: 2617–2634.

Bernheim, B., and S. Severinov. 2003. Bequests as signals: An explanation for the equal division puzzle. *Journal of Political Economy* 111: 733–764.

Bernheim, B., A. Schleifer, and L. Summers. 1985. The strategic bequest motive. *Journal of Political Economy* 93: 1045–1076.

Carnegie, A. 1962. The advantages of poverty. In *The gospel of wealth and other timely essays*, ed. E. Kirkland. Cambridge, MA: Harvard University Press.

Cooper, G. 1979. *Voluntary tax? New perspectives on sophisticated estate tax avoidance*. Washington, DC: Brookings Institution Press.

Cox, D. 1987. Motives for private income transfers. *Journal of Political Economy* 95: 508–546.

Davies, J. 1981. Uncertain lifetime, consumption, and dissaving in retirement. *Journal of Political Economy* 89: 561–577.

Gale, W., and M. Perozek. 2001. Do estate taxes reduce savings? In *Rethinking estate and gift taxation*,

ed. W. Gale, J. Hines, and J. Slemrod. Washington, DC: Brookings Institution Press.
Holtz-Eakin, D., D. Joulfaian, and H. Rosen. 1993. The Carnegie conjecture: Some empirical evidence. *Quarterly Journal of Economics* 108: 413–435.
Hurd, M. 1987. Savings of the elderly and desired bequests. *American Economic Review* 77: 298–312.
Johnson, B. and Eller, M. 2001. *Federal taxation of inheritance and wealth transfers*. Washington, DC: Statistics of Income Division, Internal Revenue Service. Online. Available at http://www.irs.gov/pub/irs-soi/inhwlttr.pdf. Accessed 7 Mar 2006.
Joulfaian, D. 1998. The federal estate and gift tax: Description, profile of taxpayers, and economic consequences. Office of Tax Analysis Paper 80. Washington, DC: US Department of Treasury.
Joulfaian, D. 2005. Choosing between gifts and bequests: How taxes affect the timing of wealth transfers. *Journal of Public Economics* 89: 2069–2091.
Joulfaian, D., and M. Wilhelm. 1994. Inheritance and labor supply. *Journal of Human Resources* 29: 1205–1234.
Kopczuk, W., and J. Slemrod. 2001. The impact of the estate tax on the wealth accumulation and avoidance behavior of donors. In *Rethinking estate and gift taxation*, ed. W. Gale, J. Hines, and J. Slemrod. Washington, DC: Brookings Institution Press.
Kotlikoff, L., and L. Summers. 1981. The role of intergenerational transfers in aggregate capital accumulation. *Journal of Political Economy* 89: 706–732.
Light, A., and K. McGarry. 2003. Why parents play favorite: Explanations for unequal bequests. *American Economic Review* 94: 1669–1681.
McGarry, K. 1999. Inter vivos transfers and intended bequests. *Journal of Public Economics* 73: 321–351.
Menchik, P. 1980. Primogeniture, equal sharing, and the U.S. distribution of wealth. *Quarterly Journal of Economics* 94: 299–316.
Modigliani, F. 1988. The role of intergenerational transfers and life cycle saving in the accumulation of wealth. *Journal of Economic Perspectives* 2(2): 15–40.
Perozek, M. 1998. A reexamination of the strategic bequest motive. *Journal of Political Economy* 106: 423–445.
Pollak, R. 2003. Gary Becker's contributions to family and household economics. *Review of Economics of the Household* 1: 111–141.
Poterba, J. 1998. Estate and gift taxes and incentives for inter vivos giving in the U. S. *Journal of Public Economics* 79: 237–264.
Tomes, N. 1981. The family, inheritance, and the intergenerational transmission of inequality. *Journal of Political Economy* 89: 928–958.
U.S. Congressional Budget Office. 2004. *The estate tax and charitable giving*. Washington, DC: Congressional Budget Office.
Wilhelm, M. 1996. Bequest behavior and the effect of heirs' earnings: Testing the altrusitic model of bequests. *American Economic Review* 86: 874–892.

Inheritance Taxes

Joseph A. Pechman

Taxes on property left by individuals to their heirs are among the oldest forms of taxation. In societies in which property is privately owned, the state protects the property rights of the individual and supervises the transfer from one generation to the next. Consequently, the state has always regarded property transfers as appropriate objects of taxation. However, taxes on bequests and gifts raise very little revenue in modern tax systems.

I

Forms of Death Taxes

Taxation of property transfers can take several forms, depending on when the transfers are made. *Estate* taxes are taxes on the privilege of transferring property to one's heirs at death. *Inheritance* taxes are levied on the privilege of inheriting property. Most estate and inheritance taxes are levied at graduated rates (sometimes reaching high levels), with high exemptions.

Taxes at death could be avoided simply by transferring property by gifts *inter vivos* (between living persons). Accordingly, estate taxes are usually associated with a *gift* tax on the donor, and inheritance taxes are associated with a gift tax on the donee. In the United States, the United Kingdom, and other countries, the estate and gift taxes have been unified into one tax. In such cases, the tax is levied on the accumulated bequests and gifts, with instalments paid on the incremental gifts as they are made and the bequest treated as the final gift.

State governments of the United States and many countries levy taxes on inheritances separately from gift taxes or without gift taxes. A unified tax on gifts and inheritances, called *accessions tax*, is not used anywhere.

Some people regard an inheritance or accessions tax as more equitable than an estate tax, because taxes are graduated according to the total wealth received by any one person. On the other hand, most countries use the estate tax because it is easier to administer.

Bequests and gifts, like income from work or investments, are a source of ability to pay. In theory, therefore, they should be taxable to the recipient as income when received. However, bequests and gifts are taxed separately from income in all countries.

History

Death taxes pre-date both income and sales taxes as a source of government revenue. The first inheritance tax was levied in the Roman Empire beginning in AD 6. During the Middle Ages, various feudal taxes resembled inheritance taxes. Such taxes were in use in several Italian commercial cities by the end of the 14th century and in England, France, Germany, the Netherlands, and Portugal by the end of the 17th century. Estate or inheritance taxes are now levied in practically all industrial countries and in many developing countries, but the type of tax and the degree of progression differs greatly among them.

England's death tax, which dates from the year 1694, took its modern form in 1779, when a flat duty was replaced essentially by a proportional tax; graduation was introduced in 1894. France adopted its inheritance tax in 1796 and introduced graduation in 1902. Italy modelled its tax after the French system in 1862 and made it progressive in 1902. Germany's tax, which was based on the Prussian inheritance tax of 1873, was graduated in 1905. The federal government of the United States levied temporary inheritance taxes during the Civil War and the Spanish–American War, but the tax was already in use in many of the states before the modern, graduated estate tax was adopted in 1916.

Rationale

Adam Smith was ambivalent about death taxes, mainly because they bear heavily on families having deaths spaced at short intervals. (This problem is now handled by providing a credit or rebate for taxes on an estate which was recently subject to tax, say, within the last five or ten years.) Smith and David Ricardo believed that death taxes would reduce the funds available for investment. Jeremy Bentham and John Stuart Mill attacked the ethical justification of the institution of inheritance, and supported a limitation on the amount any one person could acquire from others without working. Henry Sidgwick, Alfred Marshall, A.C. Pigou, and many other economists in the classical tradition supported the taxation of wealth transfers to promote greater equality of opportunity, even though some were concerned about the effect on capital accumulation. Josiah Wedgwood (1929), Hugh Dalton, James Meade (e.g. 1976) and others emphasized the objective of reducing the concentration of wealth as a justification for state or inheritance taxation. Henry Simons, who argued strongly for the taxation of what he called 'gratuitous receipts' as income, also supported a supplementary tax on wealth transfers to control the size of inheritances.

Some Keynesian economists in Britain have supported death taxes in order to raise the propensity to consume, but this rationale had no influence either on the development of these taxes or on the public's attitude toward them. Keynes himself mentioned that death taxes would probably increase the propensity to consume more than other taxes of equal yield, but did not recommend that such taxes be enacted for this reason.

Death taxes have been supported by people in all wealth classes. One of their strongest supporters was the American steel magnate, Andrew Carnegie, who had doubts about the institution of inheritance (because it impairs children's incentives) and felt that wealthy persons are morally obliged to use their fortunes for social purposes. While this view is not widely held, many people believe that the existing distribution of wealth and control of business enterprise should not be perpetuated through succeeding generations and that

taxation of bequests and gifts is the most effective method of achieving this objective. Although data are scarce, the available information suggests that inherited wealth is a major reason why the distribution of wealth is highly unequal in market economies.

Opinions about the impact of death taxes on private incentives vary. Some believe that these taxes reduce saving and undermine the economic system. But even they might concede that death taxes have less adverse effects on incentives than do income taxes of equal yield. Income taxes reduce the return from effort and risk-taking as income is earned, whereas death taxes are paid only after a lifetime of work and accumulation and are likely to be given less weight by individuals in their work, saving, and investment decisions. This distinction was emphasized by many economists in the classical tradition (including Mill, Sidgwick, Marshall, and Pigou).

Proposals have been made from time to time to tax inherited wealth more heavily than wealth accumulated out of an individual's own saving. In some plans (for example, a plan proposed by Eugenio Rignano, 1924), inherited wealth would be taxed at progressively higher rates in succeeding generations. Such proposals have never been given serious consideration because of the difficulties of tracing inherited wealth, the harshness of the tax when there are quick successions, and the problems of record-keeping and administration. It is possible to accomplish Rignano's objective by varying the tax on the basis of the number of years during which donors hold their wealth. Such a plan, which was devised by William Vickrey (1947) and later proposed in modified form by a commission headed by James E. Meade (1976), tends to be complicated (because it requires interest adjustments to equate the taxes on bequests and gifts over a given number of years) and no country has shown any interest in this type of tax.

Structural Problems

Since wealth transfers take many forms, estate and gift taxation is inherently complicated. The bases of the estate and gift taxes are intended to consist of all property transferred by gift or death, but only a small fraction of total property transfers is subject to tax. In the United States, less than one per cent of all the estates of those who die in any one year is subject to estate or gift taxes.

The estate and gift tax exemptions tend to be relatively high in most countries and concessions for transfers to children and grandchildren (consanguinity rates) are frequently excessive. The tax base is also eroded by undervaluations of farm and small business properties. Works of art and other personal property often escape taxation. In the United States, family foundations may be used to remove wealth from the estate tax base without relinquishing control of the business enterprise.

The most difficult problems in estate and gift taxation have arisen from the use of trusts to transfer wealth to future generations. Children and grandchildren may receive the income from a trust while they are living, but no tax is due on the property when the trust terminates at their death, thus avoiding tax for one or more generations. In the United States and the United Kingdom, the trust property is treated as if it is owned by the income beneficiaries, but significant tax avoidance possibilities through the use of trusts remain.

Even if the estate and gift taxes are unified, wealthy people may reduce their taxes by transferring property by gifts during their lifetimes. Usually, an annual exclusion is allowed to avoid the need to account for small gifts, but such exclusions permit large amounts of property to be transferred free of tax over a period of years. In addition, the gift tax itself is not included in the tax base, whereas the estate tax is computed on the basis of the donor's entire property, including the tax. Despite the advantage of making gifts, data for the United States suggest that wealthy people prefer to retain the bulk of their property until death.

Another device used by wealthy people to avoid estate taxes is to manipulate the ownership of various classes of stock in a corporation so as to

funnel increasing equity values to children or grandchildren. For example, by recapitalizing the equity structure of the business, the owner of a successful closely-held corporation might give his children all the common stock, which is initially given a low value, and retain the preferred stock, which is given a high value. As the corporation prospers, the common stock rises in value to reflect the increased earnings of the corporation, but this increase in wealth never shows up in the estate tax base.

Because of the practical problems of taxes on wealth transfers, some have proposed the enactment of an annual wealth tax to reach property that is not now subject to death (or income) taxes. Annual taxes on net wealth have been enacted in a number of European countries, but these taxes raise very little revenue and are regarded as supplements to estate or inheritance taxes, rather than as substitutes.

Revenue Yield

Despite the appeal of estate and gift taxes on social and economic grounds and despite the use of relatively high rates, taxes on property transfers have never provided significant revenues anywhere and have had only modest effects on the distribution of wealth. In 1983, estate and gift tax collections amounted to 0.3 per cent of gross domestic product in France, 0.2 per cent in the United States and the United Kingdom, and 0.1 per cent in Germany.

One can only guess why heavier reliance has not been placed on estate and gift taxes. One explanation is that people resent paying taxes on such wealth as the family home or business, works of art, and other personal property. The public is not aware that the major part of the estate and gift tax bases consists of stocks, bonds, and real estate, and that the exemptions remove the wealth of most people from the base. Another explanation is that greater equality in the distribution of wealth is not generally accepted as an objective of tax policy.

More intensive use of estate and gift taxes would add progressivity to tax systems with less impairment of economic incentives than many other taxes. Major obstacles to increased use of these taxes are public apathy and the lack of understanding of their major features and how they apply in individual circumstances. Resistance to higher death duties by wealthy people is also a factor. The merits of wealth transfer taxes will have to be more widely understood and accepted before they can become effective revenue sources.

See Also

- ▶ Redistribution of Income and Wealth
- ▶ Taxation of Capital
- ▶ Taxation of Wealth

Bibliography

Cooper, G. 1979. *A voluntary tax? New perspectives on sophisticated estate tax avoidance*. Washington, DC: Brookings Institution.

Meade, J.E. 1976. *The structure and reform of direct taxation*. Report of a Committee Chaired by Professor J.E. Meade. London: Institute for Fiscal Studies and Allen & Unwin.

Rignano, E. 1924. *The social significance of the inheritance tax*, Trans. W.J. Schultz. New York: Knopf.

Sandford, C.T., J.R.M. Willis, and D.J. Ironside. 1973. *An accessions tax*. London: Institute for Fiscal Studies.

Shoup, C. 1966. *Federal estate and gift taxes*. Washington, DC: The Brookings Institution.

Tait, A.A. 1967. *The taxation of personal wealth*. Urbana: University of Illinois Press.

Vickrey, W. 1947. *Agenda for progressive taxation*. New York: The Ronald Press.

Wedgwood, J. 1929. *The economics of inheritance*. London: G. Routledge & Sons.

Innis, Harold Adams (1894–1952)

Ian M. Drummond

Keywords

American Economic Association; Canada, economics in; Economic history; Innis, H. A.; McLuhan, M.; Natural resources

JEL Classifications
B31

Canadian economist, historian and university administrator. Born in rural Ontario, Innis was educated at McMaster University, Toronto (BA, MA), and at the University of Chicago (Ph.D.). Having served in the First World War, he joined the faculty of the Department of Political Economy, University of Toronto, in 1920.From 1937 until his death he was head of the department, and from 1947 he was also Dean of the Graduate School; at his death he was President of the American Economic Association.

A prolific and thoughtful scholar, Innis began by scrutinizing Canadian economic history, both in shorter writings and in such major works as *The Fur Trade in Canada* (1930) and *The Cod Fisheries* (1940), where he concentrated his attention on such great 'staple products' as codfish, fur, wheat and timber. In these works, which have been read with interest in other lands whose economic structures appear to be similar, such as Australia, Innis developed a vision of Canadian economic history that centred on the successive development of natural-resource-based industries. The physical characteristics of these industries' products, Innis believed, had shaped not only the economic but the political and cultural history of Canada. Few would now accept Innis's interpretation of Canadian history as the mere reflection of the 'staple products'. Yet for 40 years that interpretation shaped the teaching and writing of economy history in English-speaking Canada, and it affected political historiography as well. Innis's undergraduate education was aimed at the Baptist ministry, and perhaps it was a misfortune that he turned to economics; if he had followed some more speculative vocation, the particular powers of his intellect might have developed more widely and less eccentrically, although Canadian economic history would have been deprived of its most creative practitioner. The broadening of Innis's interests beyond economic history can be detected in his early writings on what he called 'the penetrative power of the price system'–the ability of market mechanisms to reshape social relationships. Uncertain in his grasp of modern economics, Innis ignored the Keynesian Revolution, and he was profoundly sceptical about the potential contribution to rational national policymaking which might come not only from economists but from other scholars; better, he thought, for university folk to concentrate upon the safeguarding of the Western cultural tradition. In his later years Innis wrote almost exclusively about very large questions – the interconnections, over very long periods, among imperial structures and means of communication. These works –*The Bias of Communications, Changing Concepts of Time, Empire and Communications, Minerva's Owl* – have had little impact on economists or economic historians, although they have influenced some students of the humanities –most notably the Canadian literary scholar Marshall McLuhan. Also, during the 1970s and 1980s Innis's writings attracted attention from Canadian nationalists, more or less regardless of discipline; furthermore, in these decades efforts were made to find and explicate new profundities in his writings, or to reinterpret this pessimistic and conservative thinker as an unconscious proto-Marxist. Few economists and fewer historians have found these efforts persuasive.

See Also

► Linkages

Selected Works

1930. *The fur trade in Canada.* New Haven: Yale University Press; revised ed., Toronto: University of Toronto Press, 1956.
1940. *The cod fisheries.* New Haven: Yale University Press; revised ed., Toronto: University of Toronto Press, 1954.
1950. *Empire and communications.* London: Oxford University Press; revised ed., Toronto: University of Toronto Press, 1972.
1951. *The bias of communications.* Toronto: University of Toronto Press.
1956. *Essays in Canadian economic history.* Toronto: University of Toronto Press.

Bibliography

Neill, R. 1972. *A new theory of value: The Canadian economics of H.A. Innis*. Toronto: University of Toronto Press.

Innovation

C. Freeman

Economists of all descriptions have accepted that new products and new processes are the main source of dynamism in capitalist development. But relatively few have stopped to examine in depth the origins of such innovations or the consequences of their adoption. Most have preferred, in Rosenberg's (1982) apt description, not to look 'inside the black box', but to leave that task to technologists and historians, preferring to concentrate their own efforts on '*ceteris paribus*' models, which relegate technical and institutional change to the role of exogenous variables.

The classical economists were generally more ready to look inside the black box; Adam Smith and Marx in particular both showed a deep interest in the relationship between scientific research, technical innovation and the market. Smith (1776), pointed already in the 18th century to the growth of specialization in scientific research and to the links between innovation in the machine-building industries and scientists ('philosophers' or men of 'speculation' whose task is 'to observe everything'). Marx and Engels (1848) probably more than any other economist assigned to technical innovation the driving force in economic development and competition – 'the bourgeoisie cannot exist without constantly revolutionizing the means of production.'

But in the first half of the 20th century Schumpeter was almost alone among leading economists in following and developing this classical tradition. Consequently those economists such as Nelson (1977, 1982) and Rosenberg (1976, 1982) who have concentrated much of their attention on the economics of innovation are often referred to as 'Schumpeterian' or 'neo-Schumpetarian', even though their ideas may considerably diverge on many topics.

It is to Schumpeter that we owe the threefold distinction between invention, innovation and diffusion of innovations, which has now become the generally accepted convention in analysis of technical change. *Invention* is generally defined as a novel idea, sketch or model for a new or improved product, process or system. It need not necessarily imply any empirical test of feasibility or prototype experience, but as Jewkes (1958) suggests, it usually does convey the first belief that something should work and often the first rough test that it will in fact work.

Nevertheless, Schumpeter was right to stress the distinction between invention and *innovation*. There is an enormous difference between 'working' under laboratory conditions and working under commercial conditions. Schumpeter used the expression 'innovation' to connote the first introduction of a new product, process, method or system into the *economy*. (This is generally taken to include military or health care applications as well as the more purely commercial innovations.) As Schumpeter pointed out, there is many a slip between cup and lip in the development of an invention to the point of commercial introduction. Problems in scaling up from laboratory scale to works scale lead to the demise of many apparently sound ideas and unanticipated 'bugs' are the rule rather than the exception in the exploitation of inventions. Many (perhaps most) inventions are patented, but most patents are never actually used commercially except perhaps as bargaining counters.

Some ambiguity still surrounds the definition of 'innovation', since the word is used both to indicate the date of *first* introduction of a new product or process (e.g. the float glass process was innovated in 1958) and to describe the whole process of taking an invention or set of inventions to the point of commercial introduction, as in 'management of innovation' – a process which may take many years of development work, trial production and marketing.

In fact the date of launch of an innovation is seldom as precise as might appear at first glance, since false starts and modifications to the design of a radical new product or process are commonplace. Thus many different dates can be found for the innovation of well-known products, such as the radio or the electronic computer. National bias plays a part too, as well as definitional problems.

This point is an important one when we come to consider the third aspect of technical change in the Schumpeterian framework – the *diffusion* of innovations. Although almost all economists would agree that the diffusion of innovations through a population of potential adopters is crucial for the achievement of productivity gains and successful competitive performance more generally, they would also agree with Rosenberg (1976), that the product or process which is being diffused is itself usually subject to further change *during* the diffusion process. Indeed, this has been one of the main criticisms of some studies of diffusion in the 1960s and 1970s (Metcalfe 1981) which tended to make the static assumptions of an unchanged product diffusing through an unchanged environment. Nevertheless, this does not invalidate Schumpeter's analytical distinction, which has proved extremely fruitful both in theoretical and empirical work, as shown notably in the major international conference on diffusion of innovations in Venice in 1986.

When Jewkes and his colleagues (1958) made their original study of the sources of invention, they rightly complained that economists had made very little contribution to the study of invention and innovation, and Rogers (1962) could legitimately make a very similar complaint, in relation to the study of *diffusion* of innovations. However, in the next quarter of a century the picture changed considerably. Following the impetus given especially by Mansfield (1968, 1977) numerous empirical studies in Europe, America and Japan covered much of the territory which Schumpeter sketched out in a preliminary way. Unknown and uncharted territory still remains, however, and its exploration is by no means straightforward (Dosi 1985).

Thus, for example, we now know a good deal about the conditions surrounding success and failure in the competitive struggle of private firms to innovate, but far less is known about the types of government policies which are most likely to encourage innovators and promote their success. The study of the latter is inhibited by the difficulty of isolating any specific single measure, such as a tax incentive, development subsidy or procurement initiative from other more general influences on the behaviour of the firm and numerous factors specific to individual firms (Rothwell and Zegveld 1981).

In the analysis of competitive attempts by individual firms to innovate the problems of multiple causality has been partly overcome by the use of statistical techniques in paired comparisons of success and failure, as for example in project SAPPHO (Freeman 1982; Rothwell et al. 1974) and similar studies in several countries (e.g. Szakasits 1974). By and large these studies agree in highlighting the main factors leading to successful innovation performance: the depth of understanding of the needs of potential users of the innovations and the steps taken to obtain this knowledge (external communications network): the research and development capability to eliminate or minimize 'bugs' prior to launch of the innovation; internal communications adequate to ensure effective links between those responsible for R&D, marketing and production within the firm; entrepreneurs or 'business innovators' with the status and experience to ensure the necessary mobilisation and coordination of resources within the firm. Studies of failure have been particularly illuminating in demonstrating the tendency of some technical innovators to neglect user needs and the lack of communication between various departments in some large firms (Burns and Stalker 1961). However, they also show that even in cases, when firms appear to follow all the 'rules' and 'best practices' which lead to good innovation performance, technical and market uncertainties may frustrate their best efforts.

Indeed, the empirical studies of the management of innovation and firm behaviour have undermined the traditional neoclassical theory of

the firm. Imperfect information, uncertainty, complex institutional linkages, cumulative in-house technology, and searching modes of behaviour are characteristic of innovation, rather than the tidy, rational, optimizing calculations and perfect foresight postulated by neoclassical theory (Dosi 1984, 1985). For this reason contributors to innovation studies have also made major new contributions to a revised theory of firm behaviour, which take into account the findings of the stream of empirical research (Nelson and Winter 1982; Dosi 1984).

Less clear-cut conclusions have emerged with respect to the influence of size and concentration on innovative performance. Schumpeter (1928, 1942) is often known for his emphasis on the advantages of large size and monopoly on innovative performance, whilst traditional theory has continued to stress the advantages of competitive market structures. Clearly large size can facilitate innovative efforts in areas where development costs are unavoidably high because of number and complexity of components, as for example in spacecraft, nuclear reactors, or electronic telephone exchanges. The R&D threshold entry barriers in such areas can sometimes be so high as to limit effective competition to only a few large organizations throughout the world; and often innovation costs are partly met by state subsidies.

Even those economists, such as Jewkes et al. (1958), who have stressed the role of individual inventors and small firms at the stage of *invention*, have accepted that often *development* costs are so high that large firms tend to predominate when it comes to *innovation*. Many of the case studies described by Jewkes et al. illustrate this point, since the small firms or individuals who initiated the inventive work were often obliged to seek the help of larger organizations or were taken over by them before they could launch the new product or process on the market.

However, revolutionary advances in technology, for example the micro-chip, can sometimes lower entry barriers dramatically. In those areas where smaller firms can afford the entry costs they appear to perform relatively well in competition with larger firms. Thus the SAPPHO project did not show size as a variable which discriminated systematically between success and failure.

Schumpeter (1912, 1928) had himself recognized the advantages of new small innovator–entrepreneurial firms, but believed that the general trend of capitalist development and the rising costs of in-house R&D would lead increasingly to the management of innovation by larger bureaucratic organizations. Galbraith (1972) developed this notion of the 'technostructure' in large firms in his 'New Industrial State'. However, empirical evidence suggest that small firms have continued to maintain, or increase their share of innovations, even though large firms do indeed now account for more than two-thirds of R&D and of all innovations (Townsend et al. 1982). The share of small firms in innovations is apparently greater than their share of R&D expenditures, and this phenomenon has been explained partly in terms of motivation and good internal communications leading to greater efficiency in the conduct of R&D, and partly in terms of the 'spin-off' of technical innovators who have left large government, industrial or academic laboratories with the idea for an already partly developed product.

The debate continues but with increasingly general acceptance that both very large and new entrepreneurial firms enjoy advantages in distinct types of invention and at different stages of the evolution of new technologies. The previously observed tendency for R&D intensity to decline in the largest firms has been denied by Soete (1979), who maintains that more recent evidence supports the Schumpeterian hypothesis.

Schumpeter's contention that technological competition was more important than price competition with invariant conditions of production has also found increasing confirmation from empirical and theoretical work in the sphere of international trade. Since Hufbauer's (1966) original demonstration of the role of technical innovation in the explanation of patterns of international trade in synthetic materials, evidence has accumulated to confirm that 'neo-technology' theories have greater explanatory power in relation to international trade performance generally than the Heckscher–Ohlin factor proportions theory (Soete 1981).

The notion that cumulative patterns of advantage in know-how, skills and innovative capability may underlie some of the persistent differences in comparative international trade and productivity performance has also found confirmation in many national studies of innovation and economic development (e.g. Pavitt 1980). These suggest that institutional innovations in education and training systems, as well as in research institutes and organizations have historically played an important part in building up cumulative technological capability. Thus, for example, German strength since the late 19th century in the chemical and engineering industries has been related to the establishment of the 'Technische Hochschulen' and other new developments in German universities, as well as the establishment of in-house R&D in the leading German chemical and electrical firms. Similar arguments have been advanced with respect to Japanese industrial training and technological innovation systems and the more recent outstanding successes of the Japanese economy (Freeman 1983).

To sum up, empirical studies of innovations and their diffusion have provided mounting evidence that mainstream neoclassical theories of firm behaviour, competition, international trade and consumer behaviour are seriously deficient in their assumptions and conclusions. However, the 'neo-Schumpeterian' tradition in economics has only begun the task of substituting a more satisfactory theoretical foundation which would take both technical innovation and institutional factors fully into account (Dosi 1985).

See Also

- ▶ Schumpeter, Joseph Alois (1883–1950)
- ▶ Technical Change

Bibliography

Burns, T., and G.M. Stalker. 1961. *The management of innovation*. London: Tavistock.

Dosi, G. 1984. *Technical change and industrial transformation: The theory and an application to the semiconductor industry*. London: Macmillan.

Dosi, G. 1985. The micro-economic sources and effects of innovation; an assessment of some recent findings. Paper given to conference on 'distribution, growth and technical progress', Rome; (mimeo) DAEST, University of Venice.

Freeman, C. 1982. *The economics of industrial innovation*, 2nd ed. Cambridge, MA/London: MIT Press/Frances Pinter.

Freeman, C. 1983. *Design and British economic performance*. London: Royal College of Art.

Galbraith, J.K. 1972. *The new industrial state*, 2nd ed. London: André Deutsch.

Hufbauer, G. 1966. *Synthetic materials and the theory of international trade*. London: Duckworth.

Jewkes, J., D. Sawers, and R. Stillerman. 1958. *The sources of invention*. London: Macmillan.

Mansfield, E. 1968. *Industrial research and technological innovation: An econometric analysis*. New York: W.W. Norton.

Mansfield, E., et al. 1972. *Research and innovation in the modern corporation*. London: Macmillan.

Marx, K., and F. Engels. 1848. *The Communist manifesto*. London.

Metcalfe, J.S. 1981. Impulse and diffusion in the study of technical change. *Futures* 13(5): 347–359.

Nelson, R.R., and S.G. Winter. 1977. In search of a useful theory of innovation. *Research Policy* 6(1): 36–76.

Nelson, R.R., and S.G. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Belknap and Harvard University Press.

Pavitt, K.L.R. 1980. *Technical innovation and British economic performance*. London: Macmillan.

Rogers, E.M. 1962. *The diffusion of innovations*. New York: Free Press.

Rosenberg, N. 1976. *Perspectives on technology*. Cambridge: Cambridge University Press.

Rosenberg, N. 1982. *Inside the black box*. Cambridge: Cambridge University Press.

Rothwell, R.R., and W. Zegveld. 1981. *Industrial innovation and public policy*. London: Frances Pinter.

Schumpeter, J.A. 1912. *The theory of economic development*. Trans. Leipzig. Cambridge, MA: Harvard University Press, 1934.

Schumpeter, J.A. 1928. The instability of capitalism. *Economic Journal* 38: 366–386.

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

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: Dent, 1910.

Soete, L.L.G. 1979. Firm size and inventive activity: The evidence reconsidered. *European Economic Review* 12: 319–340.

Soete, L.L.G. 1981. A general test of technological gap trade theory. *Weltwirtschaftliches Archiv* 117(4): 638–666.

Szakasits, G. 1974. The adoption of the SAPPHO method in the Hungarian electronics industry. *Research Policy* 3(1): 18–28.

Townsend, J., et al. 1982. Innovations in Britain since 1945. Social Policy Research Unit, Occasional paper no. 16, University of Sussex.

Input–Output Analysis

Wassily Leontief

JEL Classifications
E1

Input–output analysis is a practical extension of the classical theory of general interdependence which views the whole economy of a region, a country and even of the entire world as a single system and sets out to describe and to interpret its operation in terms of directly observable basic structural relationships.

Wassily Leontief, a Russian-born American economist, started the construction of the first input-output tables of the American economy when he joined the faculty at Harvard University in 1932. These tables, for the years 1919 and 1929, were published together with the formulation of a corresponding mathematical model and numerical computation based on it in 1936 and 1937. Thus from the very outset the new methodology – for the development of which Leontief was awarded 40 years later a Nobel prize – emphasized the importance of close mutual alignment of systematic fact finding and theoretical formulation.

In the late 1920s Leontief spent three years at the Institute for the World Economy at the University of Kiel (Germany) on derivation of statistical supply and demand curves. That early experience with curve fitting taught him not to rely on indirect statistical inference as a substitute for painstaking direct factual inquiry.

With its emphasis on disaggregation permitting detailed quantitative description of the structural properties of all component parts of a given economic system, the input–output analysis moved in a direction directly opposite to that of the highly aggregative approach that began, approximately at the same time, to dominate fundamental economic research under the powerful influence of the Keynesian paradigm presented in Keynes's *General Theory*. Hand-in-hand with a disaggregated data base went an equally disaggregated theoretical model, the empirical implementation of which involved numerical computations exceeding in their complexity and scale anything that had been carried out up to that time along these lines in economics or any other social science.

The limited capabilities of the Wilbur linear analog computer used in the first large scale computation forced Leontief to scale down his problem by neglecting some of the detail contained in the disaggregated data base. Subsequent rounds of computation were carried out at first on Howard Aiken's, Mark I and Mark II computers, and later on the early electronic machines. Thirty years later the race between the economists and statisticians compiling more and more detailed factual information, and engineers constructing more and more powerful machines, was won hands down by the latter.

A standard input–output table contains square arrays of figures arranged in chess-board fashion. Each row and the corresponding column bears the name of one particular sector, say, steel industry, automobile industry, electric power utilities, advertising services, and so on. Each individual entry represents the amount (which can, of course, be zero) of the commodity or service produced by the sector – identified by the name of the row in which it appears – that has been delivered to the sector named at the head of the column in which that entry is placed. The small schematic input–output table presented below (Table 1) describes intersectoral transactions between the three sectors of the elementary economy described by it.

Input–Output Analysis, Table 1

	Agriculture	Manufacturing	Households	Total
Agriculture	25	20	55	100 bushels
Manufacturing	14	6	30	50 yards of cloth
Households	80	180	–	260 man-years

Input–Output Analysis, Table 2

	Sector 1	Sector 2
Sector 1	0.25	0.40
Sector 2	0.14	0.12
Household	0.80	3.60

Examining these figures, one finds that to produce one bushel of wheat, agriculture requires 0.25 bushels of wheat (seed), 0.14 tons of steel and 0.80 man years of labour. A similar set of technical coefficients – 0.40 units of agricultural and 0.12 of manufactured products – describe the input requirements for production of one yard of cloth. Listed column by column these sets of technical input coefficients represent the structural matrix at the producing part of the given economy. While the figures in Table 2 were derived from the input–output table (Table 1), estimates of the magnitudes of the technical coefficients could be, and in some instances actually are, obtained directly from technical, engineering data sources.

The structural matrix of an economy provides a basis for determination of total sectoral output as well as magnitude of inter-sectoral transactions that would enable the producing sectors to deliver to households and to other so-called final users a specified 'bill of goods'. Considering the vector of final demand, consisting of 55 bushels of wheat and 30 yards of cloth, as given, the following set of balanced equations can be used to determine the total amounts of wheat (x_1), of cloth (x_2), as well as the total amount of labour (L) needed to balance under these particular technological conditions the outputs and inputs of both producing sectors,

$$\begin{aligned}(1-0.25)x_1 - 0.14x_2 \\ = y_1 - 0.40x_1 + (1-0.12)x_2 = y_2\end{aligned} \quad (1)$$

The general solution of these two equations:

$$\begin{aligned}1.457y_1 + 0.662y_2 = x_1\, 0.232y_1 + 1.242y_2 \\ = x_2\end{aligned} \quad (2)$$

permits us to compute the total levels of output of wheat, x_1 and cloth, x_2 required directly and indirectly to satisfy any given vector (y_1, y_2) of 'final demand'.

An increase in the final deliveries of agricultural products, y_1 by one unit would for instance require a rise of total agricultural output, x_1, by 1.1457 units, 0.1457 of which will have to be used to satisfy the additional input requirements of the agricultural and manufacturing sectors.

Formulated in short-hand matrix notation, the balance equations (1), describing the relationship between the column vector of final demand, y, and the column vector, x, of total outputs of all producing sectors can be written as:

$$(I - A)x = y \quad (3)$$

where A represents the upper, square part of the structural matrix describing the material input requirements of all producing sectors, x is the column vector of total outputs and y, the column vector of final deliveries of both goods. The general solution of that linear equation is,

$$x = (I - A)^{-1}y \quad (4)$$

where $(I - A)^{-1}$ represents the so-called inverse of matrix $(I - A)$. Total labour requirement can be computed in a separate step,

$$L = l'x = l'(I - A)^{-1}y \quad (5)$$

where l' is a row vector of technical labour coefficient representing the technologically determined amounts of labour that each industry employs per unit of its total output.

The same set, A, of structural coefficients that controls the physical flows, determines also the relationship between the prices of goods and services produced by different industries and the 'value added' payments (expressed in the monetary units) made by each industry per unit of its output. These include wages, profits, taxes, etc. In short, all payments other than those made for goods and services purchased from other producing sectors.

This set of value added–price equations, (often referred to as a 'dual' to set (3) of physical input–output relationships) can be formulated as follows,

I

$$(I - A')P = V \tag{6}$$

and its solution for the unknown prices as,

$$P = (I - A')^{-1}V \tag{7}$$

where P is the column vector of prices of all sectoral outputs and V is the given column vector of values added (per unit of their respective outputs), in different sectors.

In the schematic input–output table considered above all amounts entered along a particular row are measured in the same appropriately selected physical unit, for instance, wheat – in bushels; cloth – in yards; labour – in man years. No column totals are entered, since adding amounts measured in incomparable physical units would make no sense. In most published input–output tables, all transactions are measured however in value terms – usually in 'base year' prices. Since these are assumed to satisfy the price-value added equations described above – each column total, including the value added per unit of total output, must naturally be equal to the total output figures entered at the end of the corresponding row.

Value figures entered along a particular row can, however, also be interpreted as representing physical amounts of the good in question, provided the physical unit in which they are measured is implicitly defined as the quantity of that good purchasable for, say, one dollar.

In the case of a table, some rows of which are presented in conventional physical amounts, say kwh of electric power, or tons of copper, while some other rows are presented in monetary units, appropriate 'equilibrium prices' can be computed through solution of the corresponding 'dual' Eq. (7). To do so it would suffice to re-define the physical unit of the products of each sector as the amount purchasable for, say, one dollar, or some other monetary unit, at the price actually used in determination of the value figures entered on the base year table. These prices might of course be different from the equilibrium prices.

From the outset the development of input–output analysis was marked by a succession of empirical applications. In Leontief's early volume, *The Structure of American Economy, 1919–1929* (1941), this was the computation of the effects of changes in the input structure of different industries on levels of output and prices of their products, and in particular on the 'standard of living' of households.

With the onset of Second World War, attention was centred on the transition from peacetime to a war economy: in particular, on the effects of changes in the level and composition of final demand on the intersectoral distribution of output and employment. The first official US input–output table – for the year 1939, compiled for the US Bureau of Labor Statistics – provided a basis for preparation of a detailed multisectoral projection of postwar production and employment levels. Correctly predicting serious steel shortages, instead of large surpluses anticipated by leading economic and industry experts, this report gained wider interest in the new approach, not only in government circles, but among large industrial corporations as well. The Western Electric Company (the manufacturing arm of A. T.&T.) having successfully employed input-output analysis to anticipate impending shortages of lead, one of its principal raw materials, even produced an educational film describing the methodology used.

In one of the early applications of the same modelling technique as that which later on became known as operations research, the small input–output team organized – under the name Project Scoop – by the US Air Force constructed a detailed structural matrix of its far-flung material procurement and training operations. It was not a square, but rather a rectangular matrix showing for some sectors not one but several input vectors corresponding to two or more alternative technologies that could be used to produce a particular weapon or to provide a particular type of pilot training. Confronted with the problem of optimal choice between alternative 'cooking recipes', Dr. George Dantzig, a young mathematician on the Project's staff, invented the still very widely used Simplex method of linear programming, which consists of a series of inversion of structural input–output matrices with sequential substitution at alternative vectors of technical coefficients.

Not unlike research conducted in modern natural sciences, input–output analysis was from the outset most successfully conducted by closely coordinated teams rather than individual investigators. The first of such academic research groups was the Harvard Economic Research Project directed by Leontief over a period of nearly thirty years. Another centre was organized by Richard Stone in the Department of Applied Economics at the University of Cambridge. He was responsible for formal incorporation of input–output tables in the United Nations system of national accounts designed by him.

Many of the young foreign economists who came to the United States to complete or postgraduate studies spent from a few months up to several years at the HERP, and after returning home introduced input–output analysis not only as a subject of academic instruction and research but also as a new field of governmental statistics.

In Norway, Canada, Japan and in many other countries governmental planning agencies and central statistical offices compile national input–output tables and carry out practical applications of input–output analysis, but also engage in fundamental methodological research. In Soviet Russia this was the first non-marxist, mathematical approach to economics adapted, on the recommendation of Oscar Lange, after World War II as a subject of academic instruction and as a tool of economic planning.

The first International Conference on Input–Output Analysis organized by Professor Tinbergen was held in Dreibergen, Holland in 1950; the eighth has been held in Japan in 1986. Proceedings of these and of other similar scientific meetings published in book form provide a good account of the current state of the art in the general field of input–output analysis and its various applications.

One of the fundamental theoretical questions that came up in connection with the early input–output computations concerned the conditions under which none of the elements of the inverse $(I - A)^{-1}$ can be negative. The answer to it was provided by Herbert Simon – the future Nobel prizewinner – and David Hawkins, a philosopher, in the form of the following theorem:

The necessary and sufficient conditions for some of the elements of $(I - A)^{-1}$ to be positive, and all to be non-negative, are:

$$|1-\alpha_{11}|>0,\begin{vmatrix}(1-\alpha_{11}) & -\alpha_{12}\\ -\alpha_{21} & (1-\alpha_{22})\end{vmatrix} >0,\ldots\begin{vmatrix}(1-\alpha_{11}) & -\alpha_{12}\ldots & -\alpha_{1n}\\ -\alpha_{21} & (1-\alpha_{22})\ldots & -\alpha_{2n}\\ -\alpha_{n1} & -\alpha_{n2} & \ldots(1-\alpha_{nn})\end{vmatrix}>0 \tag{8}$$

If these conditions are satisfied for any particular numbering of sectors it will necessarily be satisfied for any other numbering sequence too. The economic interpretation of this theorem is that for a system, in which each sector functions by absorbing directly or indirectly outputs of some other sectors, to be able not only to sustain itself but also to make some positive deliveries to final demand, each one of the smaller and smaller subsystems contained within it has to be capable of sustaining itself and yielding a surplus deliverable to outside users as well.

An example of a system unable to sustain itself in this sense could be an economy so badly damaged by some natural catastrophe or war that only external assistance, taking the form of an import surplus, could prevent it from complete collapse. Exports are entered in a standard input-output table and in the corresponding set of balance equations, as positive and exports as negative components of the final bill of goods. The negative elements of the inverse $(I - A)^{-1}$ multiplied into such negative components of the vector y of final demand would yield in this case positive total outputs x.

In an attempt to reconcile at least to some extent the so-called fixed coefficient assumption of linear input–output models with the neoclassical production functions allowing for input substitution, Kenneth Arrow, Tjalling Koopmans and Paul Samuelson provided independently from each other three different proofs of the 'non-substitution theorem'. They considered a multisectoral economy in which each productive sector operates on the basis of a neoclassical production function and all sectors use the same single

primary factors of production, say labour. The input combinations used by different sectors are chosen so as to minimize the total amount of labour that has to be employed by that economy in order to enable it to deliver to final users an exogenously specified bill of goods. The non-substitution theorem states that the combination of the relative amounts of different inputs chosen in each sector will be independent of the composition of the final bill of goods. That means that even if the structure of final demand changes all producing sectors will behave as if they were operating on the basis of fixed coefficients of production.

Restrictive assumptions – particularly those postulating invariability of production functions that control the operations of all sectors – deprive the non- substitution theorem of much of its practical significance. However, it calls attention to the difference between the ways in which the terms technology, and technological change, are used in neoclassical and in input-output theory. In input–output modelling the technology used in any particular sector is described as a given column vector of coefficients, and a change in any element of that vector is called technological change. In neoclassical modelling the state of the technology employed by a particular sector is described by a much more general – and because of that much more complex – kind of functional relationship that in input–output analysis would have to be viewed as a set of many (strictly speaking, infinitely many) different technologies, each described by a different column vector of input coefficients. While providing a convenient basis for deductive reasoning, the neoclassical terminology makes the task of actual observation of the technological structure of a particular economy and empirical description of processes of technological change extremely, not to say prohibitively, difficult.

Since direct observation of a set of isoquants is hardly ever possible, empirical implementation of standard neoclassical models involves nearly exclusive reliance on more and more sophisticated methods of indirect statistical inference.

Neither of the two definitions of technology and technological change can be said to be more correct than the other. The employment of the simpler definition however permitted input–output analysis to advance in the direction of systematic detailed factual inquiry, while reliance on a definition, much less serviceable for purposes of empirical description but much richer in its theoretical implications, propelled neoclassical economics towards construction of elaborate theoretical models erected on a narrow, fragile data base or even on quite arbitrary, purely theoretical assumptions.

In static input–output models, additions to the stocks of building, machinery, and other kinds of productive stocks are treated as a component part of the final demand vector, entered in the right-hand side of the balance Eq. (6). In the following formulation of a simple dynamic model these terms are transferred to its left-hand sides and described explicitly as serving technologically determined capacity expansion required for a rise in the level of output.

$$(I - A)X_t - B(X_{t+1} - X_t) = Y_t \qquad (9)$$

B is a square matrix of technical capital coefficients, each column of which consists of stock-flow ratios, describing the stocks of products of different industries which the sector in question must have on hand per unit of its capacity output.

If the time unit in terms of which the process is observed and described is relatively long – say, covering a five or even ten year period – the stocks might be engaged in production in the same time period during which they have been produced. In this case, the second term on the left-hand side would be $B(X_t - X_{t-1})$. Current inputs required for maintenance of the existing capital stock have of course to be accounted for by the appropriate elements of the A matrix.

While bringing to the fore the crucial role that a complete set of capital coefficients has to play – in addition to a complete set of current input coefficients – in the detailed description of the structural framework of a given economy, such a set of difference equations is too rigid a tool to be used to describe and project the actual process of economic development and change.

More effective, because more flexible, is an approach which takes the form of a step-by-step construction of complete input–output tables of the economy for successive periods of time, each based on the knowledge of its state in the previous period, of anticipated changes in the final bill of goods and expected technological changes.

In more general terms, the input–output relationship between goods produced and consumed over a sequence of successive years can be formally described exactly in the same terms as relationships between different sectors are presented in an ordinary 'static' input-output table for a single year. The solution of a time-phased system of linear equations describing the intertemporal balances of inputs and outputs of goods and services produced and consumed over a long stretch of successive periods of time can be interpreted as inversion of a large triangular matrix; triangular because outputs of one year can become inputs in later years, but not vice versa. The results of this operation describing the direct and indirect relationships between all appropriately timed inputs and outputs has been called the 'dynamic inverse'. Since the sets of flow and capital coefficients controlling the input–output balances in successive stretches of such an historical process do not have to remain the same, both that dynamic matrix and its inverse can accurately represent all kinds of structural change, including elimination of old and introduction of entirely new goods.

Introduction of capital coefficients permits subdivision of the value-added term, V, on the right-hand side of the dual system (8) into its two parts – the returns on capital and wage income:

$$(I - A')P = \lambda B' cP + 1w \qquad (10)$$

or, solving for P:

$$(I - A')P - \lambda B' P = 1w$$

λ represents the rate of return on invested capital and w, the wage rate. These equations can be used for calculating the 'trade-off curve' between real wages (that is, money wage rate divided by a price index) and the rate of return on capital for any given state of technology. Comparison of such curves, each reflecting a different combination of alternative technologies available in different sectors, provides a base for numerical assessment of the influence of the distribution of income between the return on capital and wages upon technological choice.

Practical concerns led quite early to construction of regional input–output tables. The municipal government of the city of Stockholm was the first to compile a detailed metropolitan table. The complex fact-finding task of putting together a detailed input–output map of a particular region seemed to have been inspired sometimes by the desire to assert distinct identity. In Canada, French-speaking economists were the first to construct a regional table, that of Quebec. In Belgium one was compiled for the autonomy-seeking Flemish provinces. In addition to pressing needs of developmental planning, similar considerations seem to have prompted early compilation of input–output tables of many less developed countries.

The next step was construction of multiregional input–output tables and models in which intraregional transactions were linked with each other by interregional flows of goods and services. While comparison of labour, capital and natural resource 'contents' was the object of some of the earliest input-output studies of domestic and internationally traded goods, neither the theoretical formulation nor the available data base are yet sufficiently advanced to permit input–output modelling of international economic transactional trade to be solidly based on direct empirical implementation of the comparative cost theory. In most multiregional input–output models the structure of international transactions is controlled by sets of empirically determined export and import coefficients. A large multiregional input–output model of the world economy constructed under the auspices of the United Nations was published in 1977. Originally intended to provide a basis for a set of

alternative projections of the future growth of eight groups of developed and seven groups of less developed countries, this large, highly disaggregated model was used in a series of other studies such as the analysis of economic effects of international arms trade, detailed long-run projections of the production and consumption of non-ferrous metals in the United States and construction of alternative multiregional scenarios of future exploration of agricultural and energy resources.

As the range of its practical applications widened, the scope of input–output modelling had to be broadened, along with the contents of the requisite data bases.

Analysis of the petroleum refining industry in the early Fifties required modelling of multiproduct processes. Thirty years later a similar approach was employed to describe within the framework of a national input–output table the generation and elimination of various polluting substances. Modelling devices adapted in description of the allocation of the output of transportation and trade sectors have later on been adapted in modelling the activities of all service industries. Separation of the description of the physical from the price and costing aspects of government operations proved to be useful in construction and theoretical interpretation of input–output tables of simple, not yet fully monetized economies of the less developed economies. Richard Stone offered the conceptual framework of input-output analysis for the formal description of demographic processes.

To the extent to which it can provide a bridge between aggregative analysis and detailed description of production and consumption of specific goods and services, input–output analysis has been incorporated into most of the well-known forecasting econometric models.

The general nature of the approach has made the development of input–output analysis a cumulative process. Each refinement in theoretical structure and each addition to or improvement in the accuracy of factual information incorporated in its data base potentially improved the performance of the general model in application to all special problems.

See Also

- ▶ Hawkins–Simon Conditions
- ▶ Leontief Paradox

Bibliography

Brody, A. 1970. *Proportions, prices and planning: A mathematical restatement of the labor theory of value*. Amsterdam: North-Holland.

Brody, A., and Carter, A.P., eds. 1970. *Applications of input-output analysis*. Proceedings of the fourth international conference on input-output techniques, Geneva, 8–12 January, 1968, vol. 2. Amsterdam: North-Holland.

Brody, A., and Carter, A.P., eds. 1972. *Input-output techniques*. Proceedings of the fifth international conference on input-output techniques, Geneva, January, 1971. Amsterdam: North-Holland.

Bulmer-Thomas, V. 1982. *Input-output analysis in developing countries: Sources, methods and applications*. New York: Wiley.

Carter, A.P. 1970. *Structural change in the American economy*. Cambridge, MA: Harvard University Press.

Leontief, W. 1941. *The structure of American economy, 1919–1939: An empirical application of equilibrium analysis*. 2nd ed., enlarged. White Plains: International Arts and Sciences Press, 1951.

Leontief, W. 1966. *Input-output economics*. 2nd ed. New York: Oxford University Press, 1986.

Leontief, W., and F. Duchin. 1985. *The future impact of automation on workers*. New York: Oxford University Press.

Leontief, W., et al. 1953. *Studies in the structure of the American economy: Theoretical and empirical explorations in input-output analysis*. White Plains: International Arts and Sciences Press.

Leontief, W., A.P. Carter, and P.A. Petri. 1977. *The future of the world economy. A United Nations study*. New York: Oxford University Press.

Meyer, U. 1980. *Dynamische input-output-modelle*. Königstein: Athenaum Ökonomie Verlag.

Miller, R.E., and P.D. Blair. 1985. *Input-output analysis: Foundations and extensions*. Englewood Cliffs: Prentice-Hall.

Polenske, R., and Skolka, Jiři V., eds. 1976. *Advances in input-output analysis*. Proceedings of the sixth international conference on input-output techniques, Vienna, 22–26 April, 1974. Cambridge, MA: Ballinger.

Schumann, J. 1968. *Input-output-analyse*. Berlin and Heidelberg: Springer.

Smyshlyaev, A., ed. 1983. *Proceedings of the fourth IIASA task force meeting on input-output modeling 29 September–1 October 1983*. Laxenburg: International Institute for Applied Systems Analysis.

Inside and Outside Money

Ricardo Lagos

Abstract

A distinction is drawn between outside money, which is either of a fiat nature or backed by some asset that is not in zero net supply within the private sector, and inside money, which is an asset backed by any form of private credit that circulates as a medium of exchange.

Keywords

Banking theory; Bonds; Commitment; Fiat money; Finance theory; Inside and outside money; Money; Open market operations; Patinkin, D; Private credit

JEL Classification

D4; D10

Money is an asset that serves as a medium of exchange.

Outside money is money that is either of a fiat nature (unbacked) or backed by some asset that is not in zero net supply within the private sector of the economy. Thus, outside money is a net asset for the private sector. The qualifier 'outside' is short for '(coming from) outside the private sector'.

Inside money is an asset representing, or backed by, any form of private credit that circulates as a medium of exchange. Since it is one private agent's liability and at the same time some other agent's asset, inside money is in zero net supply within the private sector. The qualifier 'inside' is short for '(backed by debt from) inside the private sector'.

Background

In 1960, John G. Gurley and Edward S. Shaw published *Money in a Theory of Finance*, in which they attempted to develop a theory of finance that encompasses the theory of money and a theory of financial institutions that includes banking theory.

Consider a simple economy similar to the one considered by Gurley and Shaw. The economy has fiat money – an intrinsically useless asset with no backing whatsoever – that is generally accepted as a means of payment. A monetary authority or 'government' has the monopoly over issuing this asset. The economy is closed and consists of three sectors: households, firms and government. Firms issue debt in the form of homogeneous, perfectly safe nominal bonds. (For example, think of these bonds as being promises to pay one dollar at some future date.)

Table 1 shows hypothetical sectoral balance sheets for this economy. In this example, households hold only financial wealth (that is, no real wealth such as houses), in particular money, equity in firms, and the bonds issued by the firms. Here households have no liabilities, so their net worth (NW) is just the sum of the value of their assets. The assets owned by firms consist of cash and physical capital. A part of these assets has been financed with debt (bonds), and another part by issuing equity. The former represent the firms' liabilities toward the bond holders, and the latter represent the firms' liabilities towards share holders. The firms' net worth (net of equity) is

Inside and Outside Money, Table 1

Households				Firms				Government		
Assets		Liabilities		Assets		Liabilities		Assets	Liabilities	
Money	50			Money	100	Bonds	25		Money	150
Bonds	25			Capital	200	Equity	275			
Equity	275									
		NW	350			NW	0		NW	−150

zero. The government has no real assets, but at some point in the past it issued financial assets – money – to pay for expenditures, and from an accounting point of view these outstanding government-issued pieces of paper constitute liabilities. (If the money was backed by a real asset, for example gold, and also fully convertible, then the value of the gold would show up on the government's Assets column. In this case, the money issued is literally a liability representing the government's commitment to redeem the money for gold. In the case of fiat money, there need not be a counterpart on the Assets column of the government's balance sheet.)

Table 2 shows what happens if we consolidate the balance sheets of the private sector. The bonds are debts from private agents (in this example the firms) to other private agents (in this example the households), so they have cancelled out. The only assets left in the balance sheet of the public sector are physical capital and the money issued by the government. Money can be thought of as a 'claim' held by consumers and firms against the government. From the standpoint of the private sector, it is a net external, or outside, claim: it is *outside money*.

Gurley and Shaw (1960) were interested in considering the effects of 'open market operations' whereby the government issues money to purchase private bonds. Suppose, for example, that they purchase \$15 worth of private bonds. The resulting balance sheets are those in Table 3, which should be compared with those in Table 1. The government now has \$15 worth of assets (the private bonds it purchased), and its liabilities have increased by \$15 because of the money issued to pay for these bonds. Households still hold \$350 worth of assets, but the composition of their portfolio has changed: they now hold \$65 in money and \$10 in bonds, as opposed to the \$50 in money and \$25 in bonds of Table 1. The additional \$15 in money holdings comes from the new issue of money, backed by private bonds. These \$15 are government debt, but they are issued in payment for government purchases of private securities. They are a claim of consumers and firms against the world outside the private sector, but they are counterbalanced by private debt to the world outside, that is, to the government. These additional cash balances are based on internal debt, so Gurley and Shaw referred to these \$15 as *inside money*.

To use the terminology of Gurley and Shaw, the \$165 stock of money in the economy of Table 3 consists of \$150 of outside money and \$15 of inside money. Both types of money are really the same physical object, for example, green pieces of paper: The qualifiers *inside* and *outside* refer to the asset counterpart of the money. Inside money is backed by private domestic debt. Outside money is of a fiat nature (or backed by some other asset that is not in zero net supply within

Inside and Outside Money, Table 2

Combined private sector				Government		
Assets		Liabilities		Assets	Liabilities	
Money	150				Money	150
Capital	200					
		NW	350		NW	−150

Inside and Outside Money, Table 3

Households			Firms				Government			
Assets	Liabilities		Assets		Liabilities		Assets		Liabilities	
Money	65		Money	100	Bonds	25	Bonds	15	Money	165
Bonds	10		Capital	200	Equity	275				
Equity	275									
	NW	350			NW	0			NW	−150

the private sector, such as gold). Note that, if we consolidate the balance sheets of the private sector in Table 3, the net worth of the private sector is still \$350, just as in Table 2. Also, note that inside money is 'endogenous' in that if, for example, firms pay off their whole debt, *ceteris paribus* the money supply would shrink by \$15.

Most likely, Gurley and Shaw were led to stress the distinction between inside and outside money because they viewed money and private debt as assets that played distinct roles in exchange, so that an economy with the balance sheets of Table 1, where households hold \$50 in cash and \$25 in private bonds, would function differently from an economy with the balance sheets of Table 3, where households hold \$65 in cash and \$10 in private bonds. (See Gurley and Shaw 1960, pp. 82–88, the section titled 'Monetary Policy in a Modified Second Model'.) The theoretical analysis throughout the book is predominantly verbal, so it is not clear which are the precise trade-offs that agents consider when making a portfolio decision between money and bonds. The fact that households treat them as different assets is explicit in the Mathematical Appendix, where Alain C. Enthoven assumes distinct reduced-form demand functions for the two financial assets. Note that, since bonds are nominal and riskless in this set-up, it is not obvious why households would not treat them as perfect substitutes for money.)

The contemporary literature on monetary theory in general, and the subfield that deals with inside and outside money in particular, does not take it as given that money and bonds play different roles. Instead, it seeks to understand whether they indeed do, and whether they ought to. The recent emphasis has been on trying to gain a deeper understanding of the precise roles that fiat money and private debt play and ought to play, both as media of exchange and as vehicles to channel resources across economic agents, towards their most efficient use. This change of emphasis has led to a slightly different definition of inside money. The more modern use of the concept does not rely on the type of open market operations of Gurley and Shaw. Inside money need not be defined narrowly as circulating fiat money backed by private debt; the private debt *itself* is regarded as inside money if it circulates as means of payment among the private agents. The more modern definition given at the beginning of this article encompasses both the case where private debt circulates directly and Gurley and Shaw's original example. To illustrate, consider again the economy of Table 1. According to the modern use of the term, there is not enough information in that table to decide how much inside money there is in the economy; there are \$25 of inside assets, that is, assets that are in zero net supply within the private sector, but whether these assets constitute inside money depends on whether they circulate as means of payment. If they do not – for example if lenders merely hold the bonds until maturity to redeem them – then these bonds are not inside money.

Contemporary Perspectives

Gurley and Shaw (1960) simply asserted that agents would want to hold government-issued fiat money (this weakness was stressed by Patinkin 1961), and for their purposes the distinction between inside and outside money was relevant because they implicitly regarded them as imperfect substitutes. The modern literature on monetary theory seeks to identify the fundamental features of the basic economic environment that can make fiat money, or, more generally, any asset that serves as a medium of exchange, valuable and socially beneficial. Modern theory also focuses on the differences and similarities between inside and outside money. When is outside money valued? Under which circumstances does inside money arise? Are inside and outside money substitutes or complements? Under which circumstances can they coexist? Are they *both* needed to achieve efficient outcomes?

Inside money is private debt that also circulates as a tangible medium of exchange. Thus, an economy with inside money must perform a delicate balancing act. On the one hand, it must have enough commitment or enforcement for credit to be feasible, but at the same time credit must not function too well, for otherwise a tangible medium of exchange would be inessential. For

example, Kocherlakota (1998) shows that a tangible medium of exchange is not essential if agents can commit to future actions or if their trading histories are public. Starting from this observation, Cavalcanti and Wallace (1999a) consider an environment where trading histories are public for a subset of agents but private for the rest, and show that a social optimum requires note issue by those agents with public trading histories. In addition, those notes are in turn used in trade among the agents whose trading histories are private. Thus, in their environment an optimum requires inside money.

Kiyotaki and Moore (2002a) instead consider an environment where everyone is anonymous, and emphasize the importance of the agents' ability to make bilateral and multilateral commitments. The degree of (bilateral) commitment a borrower can make to an initial lender when selling a paper claim places a bound on the entire stock of private debt. The degree of (multilateral) commitment a borrower can make to repay any bearer determines the extent to which the borrower's debt can circulate in equilibrium. Kiyotaki and Moore find that only outside money circulates in economies with very low degrees of bilateral commitment. For higher, but still low, degrees of bilateral commitment, outside and inside money circulate alongside each other in equilibrium. For yet higher degrees, only inside money circulates, and, when the agents' ability to make bilateral commitments is large enough, the economy can manage without any money, inside or outside.

Acknowledgments I thank Narayana Kocherlakota and Warren Weber for comments. I also thank the C.V. Starr Center for Applied Economics at New York University for financial support. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

Bibliography

Cavalcanti, R. de, O. and Wallace, N. 1999a. A model of private bank-note issue. *Review of Economic Dynamics* 2: 104–136.

Cavalcanti, R. de O. and Wallace, N. 1999b. Inside and outside money as alternative media of exchange. *Journal of Money, Credit, and Banking* 31: 443–457.

Cavalcanti, R., A. Erosa, O. de, and T. Temzelides. 1999. Private money and reserve management in a random-matching model. *Journal of Political Economy* 107: 929–945.

Gurley, J., and E. Shaw. 1960. *Money in a theory of finance*. Washington, DC: Brookings Institution.

Kiyotaki, N. and Moore, J. 2002a. *Inside money and liquidity*. Mimeo, London School of Economics.

Kiyotaki, N. and Moore, J. 2002b. Evil is the root of all money. *American Economic Review* 92(2): 62–66.

Kiyotaki, N., and J. Moore. 2005. Financial deepening. *Journal of the European Economic Association* 3: 701–713.

Kocherlakota, N. 1998. Money is memory. *Journal of Economic Theory* 81: 232–251.

Patinkin, D. 1961. Financial intermediaries and the logical structure of monetary theory: A review article. *American Economic Review* 51(1): 95–116.

Wallace, N. 2000. Knowledge of individual histories and optimal payment arrangements. *Federal Reserve Bank of Minneapolis Quarterly Review* 24(3): 11–21.

Insider Trading

Andrew Metrick

Abstract

Insider trading has two definitions: securities trading by a corporate insider, and securities trading while in the possession of material non-public information about the security. This article reviews the two main strands of economic literature on insider trading. First, scholars on the intersection of law and economics analyse the social-welfare implications of insider-trading regulation. Second, financial economists use empirical evidence on insider trading to analyse the efficiency of stock markets.

Keywords

Adverse selection; Capital asset pricing model; Cost of capital; Efficient markets hypothesis; Informational efficiency; Insider trading; Law, economic analysis of; Market liquidity; Non-public information; Regulation costs; Risk-adjusted returns; Securities and Exchange Act 1934 (USA); Securities and Exchange

Commission (USA); Securities regulation; Short selling; Stock markets; Tipping chains

JEL Classifications
G1

Federal securities law defines a 'corporate insider' to be an officer, director, or major shareholder of a corporation. The first definition of 'insider trading' refers to any purchase or sale of public-corporation stock by an insider of that corporation. The second definition does not require the trader to be a corporate insider, but does require that the trader possess material non-public information. Within this article, all uses of the generic term 'insider trading' encompass both of these definitions. When necessary, the two definitions are referred distinctly as 'trading by insiders' (any transaction that is made by a corporate insider) and 'trading on inside information' (a transaction that requires material non-public information but need not be made by a corporate insider).

In the United States, prior to 1934 insider trading was regulated by state-level corporate law. In the first few decades of the 20th century, states used a variety of criteria to adjudicate cases, with a substantial minority of states holding that corporate directors had a duty to disclose material information before buying (but not selling) stock. Federal regulation of insider trading did not begin until the Securities and Exchange Act (SEA) of 1934. Rule 10b of the SEA made it unlawful for any person 'to use or employ, in connection with the purchase or sale of any security registered on a national securities exchange or any security not so registered, any manipulative device or contrivance in contravention of such rules and regulations as the Commission may prescribe (Bainbridge, 2001). This sweeping language does not directly mention either corporate insiders or material non-public information, but later judicial interpretations expanded its scope to these cases. Corporate insiders do appear in Section 16a of the SEA, which requires that open- market trades by insiders be reported to the Securities and Exchange Commission (SEC) within ten days after the end of month in which they took place. These reports, filed on the SEC's 'Form 4', are the source of data for almost all of the empirical studies of trading by insiders.

It was not until 1961 that the SEC took its first administrative action on an insider-trading case (Cady, Roberts, & Co.), and it would be another seven years before the first federal insider-trading case was decided by the courts *(SEC* v. *Texas Gulf Sulphur Co.* (1968)). In the decades since these seminal cases, the courts have reaffirmed and expanded the SEC's role in the regulation of trading on inside information. Despite the long judicial record, there is still considerable confusion and a continuing evolution about the scope of regulation, with debates about the type of information that is considered to be 'non-public' or 'material', and about the necessity of the trader having some fiduciary relationship to the company. A discussion of these issues is beyond the scope of this survey; readers are referred to Bainbridge (2001) for a summary.

The United States was the first country to use securities regulation to prohibit trading on inside information. Other countries were slow to adopt similar regulations: Bhattacharya and Daouk (2002) report that, as of 1990, of 103 countries with stock markets, only 34 had any prohibitions on insider trading, and only 9 had enforced their prohibitions with a prosecution. The same paper, however, reports that, by 1998, 87 countries had prohibitions and 38 had made at least one prosecution.

The remainder of this article reviews the two main strands of economic literature on insider trading. First, scholars on the intersection of law and economics analyse the social-welfare implications of insider-trading regulation. Second, financial economists use empirical evidence of trading by insiders to analyse the efficiency of stock markets. Each of these two topics has developed an extensive literature since the 1960s.

Social Welfare

Prior to the 1960s, scholars gave little thought to the social-welfare implications of insider-trading regulation. With the Cady case of 1961, the first

federal regulation in the United States stimulated a large literature on the topic, beginning with Manne (1966). The economic debate revolves around six main issues: market liquidity, informational efficiency, market manipulation, efficient managerial compensation, the costs of regulation, and the necessity of federal law.

Market liquidity. The pro-regulation side argues that, if trading on inside information were pervasive, then non-insiders would be discouraged from trading, thus reducing market liquidity and all of the other good things that come from having well-functioning capital markets. The logic here is straightforward: if non-insiders perceive that counterparties are likely to possess inside information, then they face an adverse-selection problem, and will demand a discount (if buying) or premium (if selling). The resultant spread between bid and ask prices would then act effectively as a tax on every transaction, which lowers the amount of trade.

In response to this argument, the anti-regulation side argues that the total amount of insider trading is very small, and is thus unlikely to create much of an adverse-selection problem for most stocks. Under the current regulatory regime in the United States, these adverse-selection costs do indeed seem to be low. Jeng et al. (2003) examine all reported trading by insiders in the United States from 1975 to 1996. After estimating the profits earned by insiders on these trades, the authors estimate that non-insiders have expected trading losses of about ten cents per $10,000 trade for non-insider sales and less than one cent per $10,000 trade for non-insider purchases. These results require two caveats. First, the study considers only the trades that were reported to the SEC. If the most profitable trades by insiders go unreported, then non-insiders may face larger expected losses. Second, these costs reflect the regulatory regime in place during the relevant period in the United States. If insider-trading restrictions were significantly loosened, then the frequency and profitability of insider trades might be quite different.

Informational efficiency. A second argument in favour of regulation is that, in the absence of regulation, insiders might be induced to hoard information until such time as it could be exploited in the most profitable way. For example, suppose that the managers of company XYZ have just learned of a major problem at one of their production facilities, which they expect to reduce company value by ten per cent. At the same time, managers also learn that a major research breakthrough has been made on another project, which would have an offsetting effect on firm value. Under these assumptions, if both pieces of information were immediately released, there would be no stock-price reaction. However, if insider trading were always permitted, managers would have an incentive to delay one of these announcements. For example, managers could release the bad news first, decreasing the stock price, and then buy stock in advance of releasing the good news.

The anti-regulation side provides a direct counterargument, claiming that insider trading is likely to speed up the flow of information to the market. As a counter to the example presented above, imagine that managers learn only the bad news about the production facility, with no good news about research. In this case, one can imagine these managers trying to contain this information for as long as possible, perhaps in the hope that the problem can be fixed before it is made public. In a regime without any insider trading, this strategy might be possible. With no restrictions on insider trading, however, managers would have a strong incentive to sell shares. In an extreme case, they could even sell shares they do not own ('short selling'), thus providing a virtually unlimited amount of selling, and driving the price to its 'correct' level. Opponents of regulation argue that this kind of scenario is common, and that insider trading would allow stock prices to adjust more quickly to new information. Unfortunately, there is no empirical evidence to give us more insight into this debate, nor is it easy to imagine a plausible data-set that could provide such evidence.

Market manipulation. Once again, consider the situation of company XYZ, with problems at its production facility and the potential of research breakthroughs. For managers who live through these events, it is only a short leap to imagine the possibilities of market manipulation. For

example, a well-placed rumour – coming from an insider – could move the stock price and allow for profitable trading. Opponents of regulation could counter that market manipulation can be illegal, even if trading on inside information is not. The game can grow more complex, however, if managers engage in real activities that allow for higher volatility and increased trading opportunities. For example, a CEO can increase expenditure on research and development well beyond optimal levels, safe in the knowledge that this combination of projects will increase the real underlying volatility of corporate value. In this case, the manager is manipulating the economic activities of the firm in an economically wasteful manner.

Efficient managerial compensation. Opponents of regulation argue that profitable insider trading is mostly a transfer of wealth from shareholders to managers, and thus can be treated the same as any other form of managerial compensation. For example, if shareholders believe that the CEO of their company can earn about $5 million per year from trading on inside information, then the company can reduce the CEO's other compensation by that same amount. In this scenario, the shareholders are not injured at all by the insider trading. Of course, this argument rests on the absence of the other costs discussed above: market liquidity, informational efficiency, and market manipulation.

The costs of regulation. Opponents of regulation argue that effective enforcement would be prohibitively costly. Insiders have many vehicles to exploit their superior information. In addition to stock trading in their own account, they can tip other traders, sometimes using complex 'tipping chains' that are difficult to detect. Furthermore, insiders may be able to exploit superior information by not trading at all. For example, if a manager of XYZ was planning to buy stock, but then learns bad news about a production problem, he could then decide not to buy. Since the manager has taken no action, there is no conceivable way that this exploitation of inside information could be detected. Of course, these opportunities for insider 'nontrading' are limited, since they presuppose a standing (but reversible) decision to trade.

The importance of this argument is ultimately an empirical question. In the absence of more complete information, it is impossible to know the frequency of different kinds of trading opportunities and the costs of detecting each type. Proponents of regulation can also argue that, even when detection probabilities are low, sufficiently high penalties can still provide effective deterrence.

Necessity of federal law. Prior to the Cady case of 1961, insider trading in the United States was governed by state law. Opponents of regulation argue that these state laws are sufficient, and the regulation of insider trading under federal securities laws is illogical and inefficient. There is much legal scholarship to support this view (Bainbridge, 2001), as the legal theories of insider trading are still struggling for a solid foundation, having adopted and discarded several models in the decades since Cady. The economic justification for leaving insider-trading regulation to the states rests on the identification of insider trading as a private issue between a company and its shareholders, with no externalities to security markets. If it is indeed a private issue, then opponents of regulation are correct that insider trading is the purview of other corporate law, which is left to individual states. If externalities exist – for example due to effects on market liquidity or informational efficiency – then federal regulation can be justified.

Overall, these six issues comprise the main topics of debate between the proregulation and anti-regulation sides. As seen by this survey, the empirical evidence on each of these issues is limited. For the debate as a whole, the best evidence comes from the aforementioned paper of Bhattacharya and Daouk (2002). After surveying the 103 countries with stock markets to assess the existence and enforcement of insider-trading laws, the authors used a variety of methods to estimate the cost of capital in each country. They find significant evidence that the cost of capital falls after the first enforcement of insider-trading laws. In contrast, the establishment of laws (prior to the first enforcement) has no effect on the cost of capital. Thus, for some combination of reasons – liquidity, informational efficiency, and so on – it is cheaper for firms to raise capital in

markets that enforce prohibitions against trading on inside information.

Market Efficiency

While law-and-economics scholars focused on social welfare, financial economists saw a good opportunity to use insider-trading data to test market efficiency. This literature began in earnest with the definitions of the efficient markets hypothesis (EMH) (Roberts, 1967), which comes in three versions: weak, semi-strong, and strong. Weak-form efficiency means that current asset prices incorporate all information contained in past prices; semi-strong efficiency means that current asset prices incorporate all public information; strong-form efficiency means that current asset prices incorporate all relevant information, both public and non-public.

Data on trading by insiders can be used to test both the strong and semi-strong versions of the EMH. If the strong form of the EMH holds, then insiders should not be able to make excess profits on their trades, since any information possessed by insiders would already be incorporated in market prices. One can test this implication of the EMH by analysing the risk-adjusted returns earned by insiders, where the main complication is the definition of 'risk-adjusted returns'. The capital asset pricing model (CAPM) was the first model of risk-adjusted returns to be widely adopted by economists. Finnerty (1976) uses the CAPM to evaluate the equally weighted returns to all insider trades in NYSE stocks from 1969 to 1972. He finds that insider buys overperform and insider sales underperform their CAPM benchmarks, thus providing the first direct evidence against the strong form of the EMH.

In the decades that followed Finnerty's study, researchers developed several other methods of computing risk-adjusted returns. Jeng et al. (2003) test the strong form of EMH using these more modern methods on 25 years of disclosed insider trading: they conclude that insiders earn positive risk-adjusted returns on their purchases but not on their sales. Since both the Finnerty and Jeng, Metrick and Zeckhauser studies focus on transactions reported to the SEC, they may both be underestimating insider profits if the most profitable transactions are unreported. While comprehensive data on unreported transactions is, by definition, unavailable, a unique study by Meulbroek (1992) does provide some evidence. Using proprietary data from SEC investigations of insider trading, then-SEC employee Meulbroek concluded that these transactions earned substantial risk- adjusted profits. Overall, the Finnerty, Jeng, Metrick and Zeckhauser, and Meulbroek studies provide significant evidence against the strong form of the EMH.

While the strong form of the EMH is of interest to regulators and academics, the semi-strong version commands far greater attention from investors; if the semi-strong version is false, then there exist profitable trading strategies based on public information. Economists have focused on insider-trading data as one possible source of such information. The first study of this data is Smith (1941), who finds no trading advantage for insiders, a result that discouraged other researchers until the work of Lorie and Niederhoffer (1968). These authors point out the severe problems of the SEC data, with trade dates often off by several weeks. These data problems invalidated the Smith study and opened the door to a new generation of analyses.

To handle these problems, Lorie and Niederhoffer devised a strategy that has dominated the insider-trading literature to this day: analyse the risk-adjusted returns to firms in relation to the 'intensity' of insiders' purchases and sales over well- defined periods. For example, a stock may be labelled an 'insider buy' for a month if at least three insiders bought the stock and no insiders sold it. In the decades that followed, many authors adopted this methodology, with the most important examples being Jaffe (1974) and Seyhun (1986). These many studies use a variety of intensive-trading criteria for many different sample periods, and are nearly unanimous in concluding that stocks that are intensely bought tend to outperform relevant benchmarks over a subsequent period, and that those that are intensely sold tend to underperform. They provide mixed evidence on whether other investors can profit, after

transactions costs, by using this information. Seyhun (1998) summarizes this evidence and concludes that several different trading rules lead to profits. Overall, this literature provides strong evidence against the semi-strong version of the EMH. As in all tests of the EMH, this conclusion is specific to the time period studied and the models used to estimate risk-adjusted returns. Defenders of the EMH can always propose that the effect will go away once investors learn about it, or that researchers will discover some additional risk factor to explain the results.

Conclusion

After 40 years of intense study, research in insider-trading has made substantial progress. Scholars of law and economics have identified the main arguments for and against the regulation of insider trading, and the limited empirical evidence on these arguments has sharpened the debate for future researchers. Further progress is most likely using data-sets from the many countries that have recently begun to regulate insider trading. For financial economists, the evidence on market efficiency is more straightforward. There is significant evidence that insiders profit on their own trades, and that outsiders can profit by gleaning information from the trades of insiders. Under the assumption that there is no missing risk factor that can explain these results, this evidence argues against both the strong and the semi-strong versions of the efficient markets hypothesis.

See Also

- ▶ Capital Asset Pricing Model
- ▶ Efficient Markets Hypothesis
- ▶ Law, Economic Analysis of

Bibliography

Bainbridge, S. 2001. Insider trading. In *Encyclopedia of law and economics, volume III. The regulation of contracts*, ed. B. Bouckaert and G. De Geest. Cheltenham: Edward Elgar.

Bhattacharya, U., and H. Daouk. 2002. The world price of insider trading. *Journal of Finance* 57: 75–108.

Finnerty, J.E. 1976. Insiders and market efficiency. *Journal of Finance* 31: 1141–1148.

Jaffe, J.F. 1974. Special information and insider trading. *Journal of Business* 47: 410–428.

Jeng, L.A., A. Metrick, and R.J. Zeckhauser. 2003. Estimating the returns to insider trading: A performance-evaluation perspective. *Review of Economics and Statistics* 85: 453–471.

Lorie, J.H., and V. Niederhoffer. 1968. Predictive and statistical properties of insider trading. *Journal of Law and Economics* 11: 35–53.

Manne, H. 1966. *Insider trading and the stock market*. New York: Free Press.

Meulbroek, L.K. 1992. An empirical analysis of illegal insider trading. *Journal of Finance* 47: 1661–1699.

Roberts, H. 1967. Statistical versus clinical prediction of the stock market. Unpublished manuscript, Centre for research in Security Prices, University of Chicago.

Seyhun, H.N. 1986. Insiders' profits, costs of trading and market efficiency. *Journal of Financial Economics* 16: 189–212.

Seyhun, H.N. 1998. *Investment intelligence from insider trading*. Cambridge, MA: MIT Press.

Smith, F.P. 1941. *Management trading, stock-market prices, and profits*. New Haven: Yale University Press.

Institutional Economics

Warren J. Samuels

Abstract

Institutional economics, also concerned with resource allocation and the level and distribution of aggregate income, is primarily concerned with the organization and control of the economy, that is, its power structure, which governs whose interests count. Institutionalists have a broader or deeper set of explanatory variables, including the fundamental economic role of government, the socialization of the individual, and the consequences of a business system. Thus, prices are a function of demand and supply, these a function of markets and rights, manifest in the actions of firms and governments, and the latter a matter of business control of government.

Keywords

Adams, J.; Agricultural economics; Antitrust enforcement; Association for Evolutionary Economics; Ayres, C.; Barbash, J.; Boulding, K.; Bromley, D.; Capitalism; Clark, J. M.; Commons, J.; Competition; Conspicuous consumption; Corporatism; DeGregori, T.; Dewey, J.; Dugger, W.; Economic development; Equilibrium; European Association of Evolutionary Political Economy; Evolutionism; Fusfeld, D.; Galbraith, J. Kenneth; German Historical School; Gordon, W.; Gruchy, A.; Hamilton, D.; Heterodox economics; Hobson, J.; Hodgson, G.; Holism; Industrial organization; Institutional economics; Labour economics; Law and economics; Macroeconomics; Means, G.; Methodological individualism; Mitchell, W.; Myrdal,G.; Natural resource economics; Neale, W.; Neoclassical economics; Optimality; Parsons, K.; Perlman, S.; Peterson, W.; Post Keynesian economics; Power; Preference formation; Resource allocation; Rutherford, M.; Schmid, A.; Schumpeter, J.; Social norms; Socialism; Solo, R.; Stanfield, R.; Status; Taussig, F.; Technology; Tool, M.; Trebing, H.; Value-free economics; Veblen, T.; Waller, W.; Wealth; Weber, M.; Wieser, F. von; Witte, E.

JEL Classifications

A1; B00; B52

Apart from Marxism, which has also the character of a social movement, institutional economics has become the principal school of heterodox thought in economics. Originating in and still concentrated largely, but by no means exclusively, within the United States, institutionalism has served the dual functions of providing critiques of mainstream neoclassical (and Marxian) economics and producing an alternative conception of the economy, and of doing economic research and analysis. In so doing, it has represented in part a continuation of the German and English historical traditions, including Max Weber, as well as other writers such as John Hobson.

Early Position

The place of institutional economics thus described applies principally to the post-Second World War period. During the interwar period, the picture was substantially different. For most economists, institutionalist ideas and theories were very much a part of economics. Many economists, typified by Frank William Taussig, John Maurice Clark, Friedrich von Wieser and Joseph A. Schumpeter, did not make a fundamental distinction in their own work between institutional and neoclassical economics or, if they did differentiate the two, nonetheless pursued both modes of doing economics. They could work on aspects of the problem of organization and control and on the institutional foundations of markets pretty much simultaneously with work on the theory of competition and the working of pure abstract markets, with each enriching the other. Some economists were less eclectic in their orientation. They continued the antagonism of Thorstein Veblen, on the one hand, or developed the antagonism of those suspicious of institutionalism as another form of interventionism and as largely unreceptive to the development of mathematical formalism in economic theory, on the other hand. The work of Malcolm Rutherford and others has shown a discipline largely undifferentiated in terms of institutionalism versus neoclassicism during the interwar period.

The precise relationship of heterodox institutional economics to orthodox neoclassical economics in the post-Second World War period is complicated by several considerations: the awkward sociological status of heterodoxy within the discipline; the ambivalence within institutionalism as to the relationship, some institutionalists feeling that the two schools are complementary and others that the two are mutually exclusive; and the presence within institutionalism of two different and to some extent conflicting traditions, one emanating from Thorstein Veblen and continuing through Clarence Ayres, the other starting with John R. Commons. The Veblen–Ayres tradition focuses on the progressive role of technology and the inhibitive role of institutions; the Commons tradition is less

enamoured of the imperatives of technology and approaches institutions, as modes of collective action, more neutrally; both groups accept that actual economic performance is a function, inter alia, of both technology and institutions. Notwithstanding their differences, there is a common core of institutional analysis of perhaps no greater variety of formulation than within neoclassicism or Marxism.

Relation to Mainstream Economics

Mainstream economists maintain that the central economic problems are the allocation of resources, the distribution of income, and the determination of the levels of income, output and prices. In contrast, institutional economists assert the primacy of the problem of the organization and control of the economic system, that is, its structure of power. Thus, whereas orthodox economists have a strong tendency to identify the economy solely with the market, institutional economists argue that the market is itself an institution, comprised of a host of subsidiary institutions, and interactive with other institutional complexes in society. In short, the economy is more than the market mechanism: it includes the institutions which form, structure, and operate through, or channel the operation of, the market. The fundamental institutionalist position is that it is not the market but the organizational structure of the larger economy which effectively allocates resources.

To the extent, then, that institutional and neoclassical economists study the same questions (for example, resource allocation) the institutionalists generally encompass a broader or deeper set of explanatory variables: instead of having price and resource allocation be a function of demand and supply in a purely conceptual market, these latter are in turn related to the structure of power (wealth, institutions) which help form them. Power structure in turn is related to legal rights, thence to the use of government in forming legal rights of economic significance and thereby influencing the allocation of resources, level of income, and distribution of wealth.

Institutionalists are generally less concerned with price and resource allocation per se and more with the problem of the organization and control of the economy: that is, with performance seen as specific to power (rights) structure, as well as to technology. Institutionalists are interested, for example, in the formation and role of institutions, and the interrelations between economic and legal systems and between power and belief systems.

If institutionalists insist that the economy comprises more than the market mechanism, they also object to the equilibrium and presumptive optimality modes of analysis of neoclassical economics. The search for the deterministic technical conditions of stable equilibrium, it is felt, obscures the fundamental power and choice aspects of the economy. The search for optimality, or for optimal solutions, it is also felt, is either formally empty or can be given substance only by the introduction, typically implicitly, of antecedent normative assumptions as to whose interests count, whereas in the real world such questions have to be worked out both within institutions and through contests over institutional adjustment and reformation.

Principal Ideas

The central features of institutional thought are its holism and evolutionism. Thus the further principal themes of institutional economics include the following:

1. A theory of social change, and an activist orientation towards social institutions, through focusing on both the substantive impact of institutions on economic performance and the processes of institutional change, treating institutions not as something to be taken as given but as man-made and changeable, both deliberatively and non-deliberatively.
2. A theory of social control and collective choice, or a theory of institutions, a focus on the formation and operation of institutions as both cause and consequence of the power structure and societized behaviour of

individuals and subgroups, and as the mode through which economies are organized and controlled. Instead of focusing on the mechanics of choice from within opportunity sets, a focus on the formation of opportunity sets; instead of a focus on unfettered market freedom, a focus on the total, complex pattern of freedom and control, that is, on the formation and operation of the system of control through which both actual opportunity sets and multidimensional freedom are formed.

3. A theory of the economic role of government, as a principal social process through which both itself and other institutions of economic significance are in part formed and revised. Instead of treating government, law, and the system of rights as either given and/or exogenous, these are treated as both dependent and independent, and always critical, not merely aberrational, economic variables.
4. A theory of technology, as defining and determining the relative scarcity of all resources, as a principal force in the evolution of economic structure (including the operation of institutions) and performance, and as the basis of the logic of industrialization marking the mentality as well as the practices of modern economies.
5. The fundamental principle that the real determinant of resource allocation is not the market but the organizational – institutional, power – structure of society.
6. An emphasis on facets of the value conception which transcend price, on the values represented in and given effect by the habits and customs of social life, on the pragmatic, instrumental values ensconced in the transcendental notion of the life process of man and society, and on the constructive values latent within and given effect by the working rules of law which are both the foundation and the product of the power structure of society. Included are attempts to understand the process by which values are changed, in contrast to the orthodox assumption of given values; that is, to consider within economics such questions as where the values come from, how they are tested, and how they are changed.

In amplification of these themes one finds, for example, Veblen's emphasis on status emulation as a principal force in the formation of economic behaviour, including (through conspicuous consumption and the making of invidious comparisons) the formation of consumer demands; Commons's analysis of the evolution of the fundamental legal foundations of the modern economy; John Dewey's theory of instrumental logic and social value; John Maurice Clark's analysis of the social control of business; Wesley Mitchell's emphasis on the economy as a pecuniary phenomenon; Commons's and Selig Perlman's analyses of labour unions as a mode of representing worker interests and of generating institutional change; Edwin E. Witte's, and Commons's, efforts at creating new institutions for the embodiment and protection of rising interests and for the creative resolution of social conflict and the development of a body of analysis of institutional genesis and adjustment; and, inter alia, Veblen's and Ayres's analyses of the formation of the human belief system, including that of economists, under the impact of the contest between traditional and new ways of doing things.

Apropos of the last point, institutionalists have freely pointed to the selectivity and typically implicit nature of the operative assumptions of neoclassical analysis. They insist that, by its taking institutional or power structure as given or, more typically, by its selective specification of institutions and power structure, there is a strong tendency towards selective apologetics in orthodox economics, especially in that work which is directed to the identification of 'optimal' solutions. The institutionalist solution to such problems is that of Gunnar Myrdal: to avoid the pretence of value-free economics by making all, or substantially all and certainly the operative, value premises explicit and by generating appraisals thereof.

Accordingly, institutional economists have tended to avoid recourse to methodological individualism and to abstain from puzzle-solving research in the context of models devoid of institutional embodiment and stressing equilibrium, optimality, and purely competitive markets. They have rather attended to theoretical and empirical analyses of real-world problems, such

as the operation of particular institutions, business–government relations, and the conditions of economic development. In so far as they have dealt with economic variables at fundamental conceptual levels, such as government and rights, they have at least tried to do so in both analytically credible and non-presumptive ways.

Internal Conflict

Conflict within institutionalism has largely been on two issues. One involves the putative dichotomy of technology and institutions. The other is between those who call for government planning to modify if not replace private enterprise and those who favour private enterprise but call for strong antitrust enforcement to ensure a competitive market economy.

John Kenneth Galbraith

The best-known contemporary version of the institutionalist conception of the economy has been that of John Kenneth Galbraith. Following the course laid down by Veblen, and grafting it on to a version of Keynesian economics, Galbraith explored the corporate nature and planning modes of the business system and the impact of what he considers to be technological imperatives, the social formation of individual preferences underlying demand functions, the power and continuous interaction of the state and the corporate core of the economy, the factors and forces which influence the formation of opinion and policy in the public sector, and the inevitability of resolving conflicts of interest on the basis of some conception of public purpose.

Widespread Practice

In such fields as labour economics, industrial organization, economic development, law and economics, agricultural and natural resource economics, and macroeconomics, institutionalists, through their primary attention to power structure and belief system, in the context of their overriding concerns with social change and social control, have produced understandings of economic reality quite different from those of neoclassical economists. These contributions have come through the recent work, in addition to Galbraith, of John Adams, Jack Barbash, Kenneth E. Boulding, Dan Bromley, Thomas DeGregori, William Dugger, Daniel R. Fusfeld, Wendell C. Gordon, Allan G. Gruchy, David B. Hamilton, Gardiner C. Means, Walter C. Neale, Kenneth Parsons, Wallace Peterson, A. Allan Schmid, Robert Solo, Ron Stanfield, Paul Strassmann, Marc Tool, Harry M. Trebing and William Waller, among others. Some of this work appears in the *Journal of Economic Issues,* published by the Association for Evolutionary Economics. Also in the United States, institutional economists have joined with Post Keynesian economists and with varieties of political economists to explore empirically and theoretically topics central to those fields. In Europe, Geoffrey Hodgson and others have pursued the development and application of evolution theory to the array of institutionalist topics. Some have studied the formation, use and impact of technology and others, for example, the organizational theory applicable to the corporation. The European Association of Evolutionary Political Economy has become the major forum for European institutionalists, and even for many Americans.

Altogether this work has constituted an alternative analysis of the economic system, especially of capitalism but also of socialism, and a critique of both existing economic systems and orthodox schools of economics.

See Also

- ▶ Ayres, Clarence Edwin (1891–1972)
- ▶ Clark, John Maurice (1884–1963)
- ▶ Commons, John Rogers (1862–1945)
- ▶ Evolutionary Economics
- ▶ Galbraith, John Kenneth (1908–2006)
- ▶ Heterodox Economics
- ▶ Institutionalism, Old
- ▶ Mitchell, Wesley Clair (1874–1948)
- ▶ Post Keynesian Economics

- ▶ Veblen, Thorstein Bunde (1857–1929)
- ▶ Witte, Edwin Emil (1887–1960)

Bibliography

Canterbery, E., et al. 1984. Galbraith symposium. *Journal of Post-Keynesian Economics* 7 (1).
Dorfman, J., et al. 1963. *Institutional economics*. Berkeley: University of California Press.
Gruchy, A. 1947. *Modern economic thought*. New York: Prentice-Hall.
Parker, R. 2005. *John Kenneth Galbraith: His life, his politics, his economics*. New York: Farrar, Straus and Giroux.
Sharpe, M. 1974. *John Kenneth Galbraith and the lower economics*. 2nd ed. White Plains: International Arts and Sciences Press.
Thompson, C., ed. 1967. *Institutional adjustment*. Austin: University of Texas Press.
Tool, M. 1979. *The discretionary economy*. Santa Monica: Goodyear.

Institutional Trap

Victor Polterovich

Abstract

One of the main obstacles for successful economic development is the formation of institutional traps, inefficient yet stable norms of behaviour. Domination of barter exchange, arrears, corruption and black market activities are examples of institutional traps that have hampered reforms in transition economies. Institutional traps are supported by mechanisms of coordination, learning, linkage and cultural inertia. The acceleration of economic growth, systemic crisis, the evolution of some cultural characteristics and the development of civil society may result in breaking out of institutional traps. Examples from the history of the United States and Russia are considered.

Keywords

Arrears; Barter; Civic culture; Civil society; Coordination failures; Corruption; Cultural inertia; Hysteresis; Institutional trap; Linkage effect; Lock-in; Multiple equilibria; Path dependence; Rent seeking; Reputation; Systemic crises; Transaction costs; Transformation costs; Transitional rent; Trust

JEL Classification

P3

Institutional trap is a stable but yet inefficient equilibrium in a system where agents choose a norm of behaviour (an institution) among several options. It is usually implied that multiplicity of equilibria prevails in the system, and that an institutional trap is Pareto dominated.

The concept of institutional trap is closely related to the notion of lock-in used by Arthur (1988) and North (1990); these authors showed that inefficient technical or institutional development can be self-supporting. In fact institutional traps have been studied in many papers (see for example Ickes and Ryterman 1992; Tirole 1996; Bicchieri and Rovelli 1995; Jonson et al. 1997; Uribe 1997). In Polterovich (2000, 2005) a general scheme for the formation of an institutional trap was described. The theory developed was successful in explaining a number of important features of wide-scale institutional transformation in Russia and other post-communist countries where the evolution of institutional traps was clearly observable. In particular, it was shown that such different phenomena as barter, mutual arrears, tax evasion, and corruption were intensified and supported during the reforms due to similar mechanisms. Also studied were possible strategies for a country to get out of an institutional trap.

Norm-Fixing Mechanisms and Institutional Trap Formation

A norm is a rule that large groups of people can or must obey. In any area of life and at each moment in time, a multitude of alternative norms is available, and every agent has to make his or her choice. For example, an official may choose either corruption or honest service.

Each agent who interacts with partners within the framework of a certain behavioural norm has to bear the corresponding transaction costs. For example, the possibility of being caught while taking a bribe would cause a transaction cost component for an official who has chosen corruption as the norm.

The costs of transition from one norm to another are called transformation costs. These may be incurred by an individual, a firm or the state. If a firm decides to switch from black market to legal operations, it has to search for new partners. Search expenditure is a part of the transformation cost.

For a behavioural norm to be stable, individuals should feel that it is unprofitable or disadvantageous for them to deviate from it. This means that the present value of the difference between the transaction cost of a prevailing norm and any alternative norms has to be less than the related transformation cost. The main type of stabilizing mechanism is based on the coordination effect, according to which the more consistently a norm is observed in a society the greater are the costs incurred by each individual deviating from it. For example, the coordination effect takes place if a personal probability to be punished for a rule-breaking activity decreases with the number of people involved in the activity. In this respect, institutional traps belong to a broader class of coordination failures (Howitt 2003; see also poverty traps).

With time, the transaction costs of a norm's observance decreases due to learning effect since the agents learn to operate more efficiently. If the payment of taxes is considered a norm within a society, the taxpaying technology improves. If, on the contrary, tax evasion is a norm, the relevant techniques develop. A decrease of the transaction costs fixes the norm.

Another mechanism, referred to as the linkage effect, is also important. With time, an established norm finds itself linked with a multitude of other rules, and becomes part of a system of other norms. Therefore, non-observance of this norm triggers a chain of other transformations and, consequently, leads to high transformation costs. By increasing transformation costs, the linkage effect, too, contributes to a norm's fixation.

There is yet another norm-fixing mechanism, cultural inertia, which denotes agents' reluctance to review those behavioural stereotypes that have already proven viable. Inertia effects may be supported by a formal or informal system of punishments and awards for past behaviour. For example, a person with a good reputation tries to maintain that reputation by following respectable norms of conduct.

As with any other norm, an institutional trap's stability means that a system absorbing a small external impact will remain in the institutional trap, having perhaps slightly changed its parameters, and will return to the former equilibrium state once the source of destabilizing pressure is removed. An individual or a small group of people loses if it deviates from an institutional trap. However, the simultaneous adoption by all agents of an alternative norm may be Pareto improving. Thus the lack of coordination is the main cause of the institutional trap stability.

The emergence of institutional traps is an important source of risk associated with any reform process. The universal norm-fixing mechanisms described above, the coordination, learning and linkage effects, as well as cultural inertia, are responsible for institutional trap formation.

Consider a system with multiplicity of equilibria, and let an efficient norm prevail. Under a strong perturbation, the equilibrium may lose its stability or disappear so that the system moves to an alternative stable equilibrium, a potential institutional trap. After the disturbing factor is removed the system remains in the new equilibrium, which is now inefficient. This is the so-called hysteresis effect, which is a form of a system's dependence on its former path of development (path dependence).

A number of unexpected phenomena observed during the wide-scale reforms of the 1990s, including the rise and persistence of arrears, corruption, black market activity, and barter exchange, may be considered as institutional traps. Using the Russian experience, one can describe barter and corruption traps formation in greater detail.

Example 1: Barter

In modern economies, barter is associated with higher transaction costs than monetary transactions. When the inflation rate increases, paper money loses its value. Economic agents try to diminish their losses and seek to accelerate the rates of money circulation, which means an increase of their transaction costs. The transaction costs of monetary exchanges may grow very rapidly, if the finance system fails to cope with the rocketing number of transactions.

In economies with advanced banking systems the share of barter is rather modest, even when inflation is high. But after price liberalization in 1992, Russia proved to be ripe for barter. With the banking system still unformed, money transfers within Moscow could take up to two weeks, and beyond the capital, over a month. It sometimes made more sense to carry bags of cash from city to city by plane than to transfer money from one bank account to another. Many firms soon found that barter transaction costs were lower than those for monetary exchange. Moreover, the transformation costs of a shift to barter looked acceptable, given the pre-reform direct links between supplier and consumer that had been typical in the centrally planned economy. The search for prospective partners and the process of trade negotiations were facilitated by the spread of sophisticated means of communication. The larger the number of firms choosing barter, the lower the barter transaction costs for a fixed barter volume since it was easier to find partners and put together barter chains (a coordination effect). In those conditions, as the share of barter exchanges increased, even more companies became involved.

Thus the environment conducive to barter had been created by changes in fundamental parameters, such as the rate of inflation and the risk of arrears, which radically increased the ratio of monetary exchange transaction costs to barter exchange transaction costs. The coordination effect triggered a rapid formation of a barter economy. Later, the transaction costs of barter exchanges continued to decrease due to the learning effect: companies learned to design elaborate chains of barter exchanges. The newly established norm gave birth to a new institute of barter exchange intermediaries and proved to be an efficient instrument of tax evasion (linkage effect).

By 1997, inflation in Russia had decreased dramatically, and monetary exchange technology had notably improved. Barter practices, however, were not dropped altogether. Barter-driven behaviour was supported by the coordination effect; it has been fixed through learning, linkage and cultural inertia. Any agent deciding to break out of the barter system would be exposed to inevitable transformation costs. He or she would be forced to sever long-established connections, to look for new partners, and to be ready to come face to face with the tax-collecting authorities. The barter intermediaries, who would lose their main sources of income if barter practices were eliminated, formed a potential group of pressure for perpetuation of the relevant norm. This is the hysteresis effect mentioned above.

Example 2: Corruption

Every potential bribe-taker makes decisions comparing his or her gains from bribes and from honest behaviour. In Russia, income inequality jumped sharply during transition because of uneven transitional rent expropriation. The state was not able to properly adjust the salaries of bureaucrats, so the salaries were insignificant in comparison to bribes from the newly rich. This caused an increase in corruption activity. Inefficient government policy, inadequate legislation, unclear norms for new market behaviour and weak mechanisms of government control contributed to a rise in corruption.

The larger the scale of corruption, the smaller were the chances for a bribe-taker to be caught. Corruption technologies were developed with time, corruption hierarchies arose, and corruption activities were closely linked with other shadow economy mechanisms. Corruption turned out to be habitual for both the bureaucrats and the population. The coordination, learning, and linkage mechanisms as well as cultural inertia made the corruption system even more stable.

One can find institutional traps in the history of many developed countries. The United States of nineteenth century presents a good illustration of the corruption trap (Knott and Miller 1987, pp. 15–31). The time between 1815 and 1840 was a period of intensive transformations of political institutions in the United States. Property ownership requirements were abandoned to allow the lower classes to vote. These democratic reforms had unanticipated consequences, however. The political party machine became an effective instrument for some party bosses to get rich. Such men allocated public service positions (including those of postmaster, customs official, and policemen) among their supporters without taking into account competence or skills. Office workers were forced to pay a proportion of their wage to the political party through whom they had obtained their jobs. The police were a political tool rather than a law enforcement agency. Businessmen paid bribes for franchises. Low-level policemen took payments for 'permitting' local vice operations, and the money was distributed among the police hierarchy and the political bosses. Many people understood that the situation had to be changed, but nobody wanted to make a move. This was a corruption trap.

Once it has fallen into an institutional trap, the system chooses a non-efficient path of development, and, with time, returning to efficient development may be very difficult even if possible.

Escaping from an Institutional Trap

However, there are reasons to believe that some institutional traps are stable in the medium run only and that an economy can gradually develop mechanisms conducive to its escaping from institutional traps. The theory outlined above gives us a framework for the systematic consideration and classification of different mechanisms that may facilitate this transformation.

One has to reach at least one of the following goals: (*a*) to increase the transaction costs of the prevalent inefficient norm; (*b*) to decrease the transaction costs of an alternative efficient norm; (*c*) to bring down the transformation costs of the transition to an efficient norm. The coordination, linkage or/and inertia mechanisms have to be influenced for these purposes.

Below we consider microeconomic measures and macroeconomic policies that may be taken by a government, as well as spontaneous tendencies that are helpful for an economy to escape institutional traps.

Microeconomic Measures and Macroeconomic Policies

The simplest way of increasing the transaction costs of an inefficient norm is the introduction of a high penalty for deviating behaviour: for example, a strong punishment for corruption or a special tax on barter exchange. However, high penalties are very costly. There are at least three sources of penalty costs. First, enforcement of stronger penalties requires larger resources to be spent. Large fees may result in strong resistance on the part of the penalized persons. Second, a penalty directed to decrease the intensity of an inefficient norm may increase the intensity of its even more inefficient substitutes. Fee increasing may shift the system to another institutional trap instead of shifting it to an efficient equilibrium. For example, strong punishment for arrears could create additional incentives for firms to escape into the underground economy. Third, one should take into account the possibility of wrong decisions. The stronger the punishment of an innocent person, the larger the social losses.

The development of reputation mechanisms is another way of increasing the transaction costs of corruption, arrears, or tax evasion (Tirole 1996). These mechanisms also decrease transaction costs of efficient norms, creating incentives to observe them. At the start of the Russian transition, old reputation mechanisms were totally destroyed. New mechanisms arose gradually, due to strengthening of the state and formation of new business networks.

Amnesty is an instrument of weakening inertia effects in the cases of tax evasion, arrears and corruption. Many governments use this measure. The outcome is mixed, however. To be successful the amnesty has to be an unexpected event,

conducted at an appropriate moment when fundamental causes for a trap are exhausted, and it has to be complemented by other measures weakening linkage and coordination effects. The rotation of officials may be an effective measure for destroying unproductive coordination (see a theory of rotation in Ickes and Samuelson 1987).

Macroeconomic policy also influences the evolution of institutional traps. In choosing tax, social, or industrial policies, one has to take into account that they can create incentives or disincentives for participation in black market operations or corruption.

Spontaneous Exit

There are some spontaneous tendencies which, being unintended, may nevertheless facilitate exit from institutional traps.

A number of institutional traps (corruption and tax evasion traps, for example) are connected with rent-seeking behaviour. Each economic agent may invest his or her money and time into production or into rent-seeking activity. The choice depends on the relative efficiency of these two options. If rent-seeking dominates, then many agents choose this option, and an institutional trap may arise.

At a time of major institutional transformation, some economic agents are able to derive additional income – transitional rent – exclusively from their fortunate positions. Price liberalization gives the advantage to suppliers of goods in high demand. Foreign trade liberalization allows importers and exporters to profit from differences between domestic and world prices. The emergence of new stock exchanges and securities markets creates ample arbitrage opportunities for financial intermediaries.

If the state does not take special measures to extract transitional rent, rent-seeking becomes much more profitable than production. An increasing number of economic agents find themselves to be involved in rent-seeking activity, and increasing volumes of resources are diverted from productive activities. The rate of production growth falls, and this makes production even less attractive for investors. Coordination, learning, linkage, and inertia mechanisms start to work and form institutional traps.

If, however, the rate of economic growth substantially increases due to improvements of technology or term of trade, then some agents may decide to increase their investment into production. This supports growth and creates new incentives for the next cohort of agents to switch their efforts from rent-seeking to production. As a result, an institutional trap may disappear. Growth diminishes the transaction costs of 'good behaviour' and facilitates improvement of institutions. This conclusion was corroborated by econometric calculations (Chong and Calderon 2000) as well as theoretical research (Balatsky 2002).

Evolution of Civic Culture

One way out of an institutional trap is disadvantageous for each isolated economic agent but advantageous for society as a whole. The root of the problem is lack of coordination. The ability of agents to coordinate their efforts depends on the prevailing civic culture and the development of civil society.

Most studies of economic growth consider civic culture as a fixed and non-changing factor. However, some important parameters of civic culture may change drastically during a period of 10–20 years; therefore long-term considerations have to take them into account. For example, the proportion of people who revealed political interest in Germany was 27% in 1952 and 50% in 1977; the proportion of affirmative answers on the question 'Can most people be trusted?' increased from 9% in 1948 to 39% in 1976 (Conradt 1989). Political interest and social trust are important preconditions for social activity and the strengthening of civil society. Note that the proportion of respondents who belonged to a voluntary organization grew in Germany from 44% in 1959 to 50% in 1967, and 59% in 1975.

Lack of trust has direct economic consequence: it increases transaction costs and decreases investment (Zak and Knack 2001). If social activity is intensified and the degree of social trust increases, coordination becomes less costly; and there are more chances to escape from institutional traps.

The history of the US corruption trap, mentioned above, demonstrates the importance of the development of civil society (Knott and Miller 1987, pp. 33–53). By the turn of the nineteenth century, a powerful progressive movement had emerged. The movement combined the efforts of several groups of citizens including middle-class taxpayers, small businessmen, farmers, and professionals of various sorts. Their main goal was an administrative reform that would separate politics from administration. They required administration according to rules, the selection of civil officers according to merit and qualification, the standardization and simplification of procedures, the centralization of administrative authority under a single executive in accordance with the principles of hierarchy. Progressives created a number of organizations such as the New York Municipal Research Bureau, New York Citizen's Union, and the Milwaukee Free Press, and occupied leading positions in both Republican and Democratic Parties. The US Republican President Theodore Roosevelt and Democratic President Woodrow Wilson conducted reforms in accordance with progressive ideas and constructed a new system of governance based on independent commissions. The elimination of the corruption trap was a result of these reforms.

Systemic Crises

Sometimes systemic crises can be helpful in helping an economy escape from an institutional trap. (The idea that a systemic crisis may be advantageous has been put forward and studied in a number of papers: see Drazen and Grilli 1993.)

A crisis drastically changes system parameters and even destroys supporting mechanisms so that an economy may find itself outside the attraction area of the inefficient norm. The evolution of the barter trap in Russia serves as a remarkable illustration of this statement.

The barter trap was broken in 1998 due to systemic financial crisis. In consequence of the rouble devaluation the dollar has strengthened against the rouble by about two times in real term. Imports dropped drastically – in 1999 to 56% of the 1997 level. Exports decreased because of the rise in oil prices. Real wage rates also dropped. However, the overall demand for domestic goods increased, labour costs diminished and the economy started to grow. The crisis totally destroyed the government bond market, which diverted money flows from production purposes. Enterprises started to earn money and used it for investments. Their real balances increased. All these changes contributed into a strong decrease in monetary exchange transaction costs. The share of barter in industrial sales fell dramatically. In 2002 it was about 10%. The barter trap disappeared, including the complicated system of barter intermediaries. The crisis achieved what the government had not been able to do.

Conclusion

Institutional traps are serious obstacles to economic development. Many countries have found themselves in institutional traps. Some were able to escape, others have been searching for an exit for a long time.

The main cause of institutional traps is lack of coordination. The market is a powerful coordination mechanism; however, if the market fails, the government may try to prevent an institutional trap or facilitate getting out of it by developing reputation mechanisms, implementing an amnesty, improving administration and choosing appropriate macroeconomic policies. In many cases, however, neither market nor government measures are effective in the short run. Civil society institutions have to be developed to reach the necessary coordination. This is a point that may be helpful in integrating cultural and civil society studies into the theory of economic development.

See Also

- Arrears
- Barter
- Path Dependence
- Poverty Traps
- State Capture and Corruption in Transition Economies

Bibliography

Arthur, W.B. 1988. Self-reinforcing mechanisms in economics. In *The economy as an evolving complex system*, ed. P.W. Anderson, K. Arrow, and D. Pines. Santa Fe: Addison-Wesley.

Balatsky. 2002. Functional properties of institutional traps. Economics and mathematical methods (Funkzional'nye svoistva instituzional'nyh lovushek). *Ekonomika i matematicheskie metody* 38: 54–72 (In Russian).

Bicchieri, C., and C. Rovelli. 1995. Evolution and revolution: The dynamics of corruption. *Ration Soc* 7: 201–224.

Chong, A., and C. Calderon. 2000. Causality and feedback between institutional measures and economic growth. *Econ Polit* 12: 201–224.

Conradt, D.P. 1989. Changing German political culture. In *Civic culture revisited*, ed. G.A. Almond and S. Verba. Newbury Park: Sage.

Drazen, A., and V. Grilli. 1993. The benefit of crises for economic reforms. *Am Econ Rev* 83: 598–607.

Howitt, P. 2003. Coordination failures. In *An encyclopedia of macroeconomics*, ed. B. Snowdon and H.R. Vane. Cheltenham: Edward Elgar.

Ickes, B.W., and R. Ryterman. 1992. The interenterprise arrears crisis in Russia. *Post-Soviet Affairs* 8: 331–361.

Ickes, B.W., and L. Samuelson. 1987. Job transfers and incentives in complex economic organizations: Thwarting the Ratchet effect. *Rand Journal of Economics* 18: 275–286.

Jonson, S., D. Kaufman, and A. Shleifer. 1997. The unofficial economy in transition. *Brook Pap Econ Act* 2: 159–239.

North, D. 1990. *Institutions, institutional change and economic performance*. Cambridge, UK: Cambridge University Press.

Polterovich, V. 2000. Institutional traps. In *The New Russia: Transition gone awry*, ed. L.R. Klein and M. Pomer. Stanford: Stanford University Press.

Polterovich, V. 2005. Institutional traps: Is there a way out? *Social Sciences, A Quarterly Journal of the Russian Academy of Sciences* 36: 30–40.

Tirole, J.A. 1996. A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *Review of Economic Studies* 63: 1–22.

Uribe, M. 1997. Hysteresis in a simple model of currency substitution. *J Monet Econ* 40: 185–202.

Zak, P.J., and S. Knack. 2001. Trust and growth. *Economic Journal* 111: 295–321.

Institutionalism, Old

Malcolm Rutherford

Abstract

What is now called 'old' institutional economics was a central part of a pluralistic American economics during the inter-war period. It is a tradition that still exists today but as a marginal heterodoxy to a dominant neoclassical mainstream. By the early 1920s it had established itself as an appealing programme with a major presence at leading universities and research institutes. Institutionalist work over the inter-war period included significant contributions to economic measurement and analysis. A number of factors led to the decline of institutional economics after the Second World War, but institutionalism has continued in a modified form, and still attracts adherents today.

Keywords

Adams, H.C; Administered pricing; Ayres, C.E; Business cycles; Clark, J.M; Collective bargaining; Commons, J.R; Conciliation and mediation; Conspicuous consumption; Cooley, C.H; Dewey, J; Ely, R.T; Hamilton, W.H; Hobson, J.A; Household economics; Household production; Institutional economics; Institutionalism, old; Marginal utility theory; Market failure; Mitchell, W.C; Monopoly; National Bureau of Economic Research; New Deal; Overhead costs; Perfect competition; Planning; Public utilities; Public utility regulation; Separation of ownership and control; Social norms; Underconsumptionism; Veblen, T

JEL Classification
B5

What is now often referred to as the 'old' institutional economics was a central part of American economics during the inter-war period, and is a tradition of economics that still exists today.

The explicit identification of something called the 'institutional approach' to economics, or 'institutional economics', goes back to 1918 and to Walton Hamilton's American Economic Association (AEA) conference paper, 'The Institutional Approach to Economic Theory' (Hamilton 1919).

Hamilton's paper was a call for the profession at large to adopt the 'institutional approach'. For Hamilton, anything that 'aspired to the name of economic theory' had to be (i) capable of giving unity to economic investigations of many different areas; (ii) relevant to the problem of social control; (iii) relate to institutions as both the 'changeable elements of economic life and the agencies through which they are to be directed'; (iv) concerned with 'process' in the form of institutional change and development; and (v) based on an acceptable theory of human behaviour, in harmony with the 'conclusions of modern social psychology'. According to Hamilton, only an approach to economics that focused on the institutions that make up the 'economic order' could meet these tests. He identified H.C. Adams, Charles Horton Cooley (his own teachers at Michigan), Thorstein Veblen and Wesley Mitchell as the leaders of this movement. At the same session of the AEA conference, J.M. Clark (Clark 1919), argued for an economics both 'relevant to the issues of its time' and based on an 'ideal of scientific impartiality'. Walter Stewart (Hamilton's friend and colleague) chaired the session, and argued that economics needed to be 'organized around the central problem of control', should utilize the 'most competent thought in the related sciences of psychology and sociology', and combine 'the statistical method and the institutional approach' (Stewart 1919, p. 319).

The exact timing of this effort to promote 'institutional economics' as a distinctive approach probably had much to do with the end of the First World War. The war had impressed upon many the great importance of improved economic data and policy analysis, and of the potential role of government in the economy. The period of reconstruction seemed to offer significant opportunities for bringing changes to the conduct of economic research, education, and policy. The 1918 session of the AEA conference was followed by further efforts to promote institutional economics. Another AEA session critical of traditional theory was organized in 1920. This featured J.M. Clark's paper 'Soundings in Non-Euclidian Economics' (Clark 1921), which criticized orthodox theoretical propositions. In 1924 Mitchell argued in his presidential address to the AEA that quantitative methods would transform economics by displacing traditional theory and leading to a much greater stress on institutions (Mitchell 1925). Lionel Edie called this address 'a genuine manifesto of quantitative and institutional economics', one that stated 'the faith of a very large part of the younger generation of economists' (Edie 1927, p. 417). In the same year Rexford Tugwell edited *The Trend of Economics*, a book again seen as something of an institutionalist manifesto and which included papers from Mitchell and Clark as well as from younger people of institutionalist persuasion such as Tugwell himself, F.C. Mills, Sumner Slichter, Morris Copeland, and Robert Hale (Tugwell 1924).

During the inter-war period institutionalism developed a significant following, with a concentrated presence at a number of major schools and research institutes. In addition to Veblen, Hamilton, Clark, Mitchell, and Commons, who were the most visible proponents of institutionalism, there were many others associated with the movement (Rutherford 2000a, b). The two major centres for institutionalism over the whole inter-war period were Columbia and Wisconsin, at that time among the leading doctoral departments of economics in the country. Wisconsin's department included Commons (until he retired in 1933), E.E. Witte, Harold Groves, Martin Glaeser, Selig Perlman and several others (Rutherford 2006). Columbia was an even bigger centre for institutionalism with Mitchell, Clark, Rexford Tugwell, Mills,

A.R. Burns, Joseph Dorfman, Leo Wolman, Carter Goodrich, James Bonbright, and Robert Hale all in the Economics Department or Business School at various times, and Gardiner Means, Adolf A. Berle, and many other people of related views in other departments (Rutherford 2004). Chicago had an institutionalist contingent at least until Clark left for Columbia in 1926, and Walton Hamilton was at the centre of groups first at Amherst (1915–1923) and later at the Robert Brookings Graduate School (1923–1928). Other institutionalist groups existed at Texas, where Clarence Ayres joined Robert Montgomery in 1930, and in a number of other schools and colleges (Rutherford 2003, 2007).

Among research institutes, the Institute of Economics, which became part of the Brookings Institution, was heavily institutionalist in character (the research staff included Isador Lubin, and Edwin Nourse among others). The National Bureau of Economic Research (NBER) was closely associated with Mitchell's quantitative approach and his programme of business cycle research and employed many of his Columbia colleagues and students. The quantitative and policy orientation of the work done by these organizations attracted funding from foundations such as Carnegie and Rockefeller (Rutherford 2005a).

The Sources and Appeal of Institutional Economics

The elements that went to make up the core of the institutional approach as defined by Hamilton, were all present in American economics before 1918. Institutionalism as it formed in the inter-war period was an approach to economics that derived from several sources. While the single most significant source of inspiration for institutionalism was the work of Thorstein Veblen, it is important to understand that institutionalism was a blending of ideas taken from Veblen with those from others (Rutherford 2001), and was never simply Veblenism.

At the most basic level the most important element in the institutionalist approach is the conception of the economic system as a set of evolving social institutions. In this, institutions are seen as much more than constraints on individual action. Social norms, conventions, laws, and common practices embody generally accepted ways of thinking and behaving, and they work to mould the preferences and values of individuals brought up under their sway. A good part of this orientation came from Veblen, but also from sociologists such as Charles Horton Cooley, and from a previous generation of German-influenced scholars (such as R.T. Ely and H.C. Adams). At this time, in line with the German model, sociology was commonly taught within economics departments.

On a more specific level, Veblen's framework, which stressed the role of new technology in bringing about institutional change (by changing the underlying ways of living and thinking) and the predominantly 'pecuniary' character of the existing set of American institutions, was widely influential among institutionalists. Within this framework Veblen developed his analyses of 'conspicuous consumption'; the effect of corporate finance on the ownership and control of firms; business and financial strategies for profit-making, salesmanship and advertising; the emergence of a specialist managerial class; business fluctuations; and many other topics (Veblen 1899, 1904).

For Veblen, the existing legal and social institutions of America were outmoded and inadequate for the task of the social control of modern large-scale industry. Veblen perceived a *systemic* failure of 'business' institutions to channel private economic activity in ways consistent with the public interest. He attacked the manipulative, restrictive, and unproductive tactics used by business to generate income (including consolidations, control via holding companies and interlocking directorates, financial manipulation, insider dealing, sharp practices, and unscrupulous salesmanship), the 'waste' generated by monopoly restriction, unemployment, conspicuous consumption, and competitive advertising, and he held out little hope of change short of a complete rejection of 'business' principles.

Cooley also analysed pecuniary institutions but in more measured tones, and it must be

emphasized that many institutionalists, including Hamilton, Clark, Commons, and Hale placed a much greater emphasis on the evolution of legal institutions than did Veblen. Both Hamilton and Hale moved into law schools and had close connections with legal scholars of the realist school. The major sources of this emphasis on legal institutions were Ely (who taught Commons) and Adams (who taught Hamilton). This greater emphasis on law and on legal evolution helped to shift the character of institutionalism away from Veblen's radicalism and connect it to a pragmatic philosophy, based primarily on the work of John Dewey, which looked to legislative and legal reform concerning such issues such as business regulation, labour law, collective bargaining, health and safety regulations, and consumer protection. Thus, in the hands of institutionalists such as Hamilton, Clark, Mitchell, and Commons, the problem became one of supplementing the market with other forms of 'social control' of business.

Another important element was the linking of institutional economics with 'modern psychology'. Veblen had provided a particularly penetrating criticism of the hedonistic psychology implicit in marginal utility theory (Veblen 1898) and pointed to an alternative based on instinct/habit psychology. What was important for institutionalists, however, was less Veblen's specific formulation but the impetus he gave to the idea that economics needed to be reconstructed on the basis of a theory of human behaviour in harmony with the conclusions of modern psychology (see Mitchell 1910a, b).

Finally, and of central importance to the attraction of institutionalism, was the claim that it represented the ideal of empirical science. An important influence here was Mitchell's combination of Veblenian ideas concerning the significance of the institutions of the 'money economy' with the quantitative and statistical approach he had absorbed as a student at Chicago. Mitchell's *Business Cycles* (1913) was enthusiastically received and widely regarded at the time as a paradigm for a scientific economics. Mitchell thought of business cycles as a phenomenon arising out of the patterns of behaviour generated by the institutions of a developed money economy (Mitchell 1927), and he explicitly connected quantitative work and the institutional approach, arguing that it is institutions that create the regularities in the behaviour of the mass of people that quantitative work analyses (Mitchell 1924, 1925). Mitchell's quantitative bent was shared by many other institutionalists, but the scientific method, for institutionalists, was not confined to the statistical or quantitative, and included all work that was genuinely 'investigative' in character. It is important to comprehend that at this time it was institutionalists, not neoclassicals, who were claiming to be following the methods of natural science (Rutherford 1999), and seemed to be at one with the general movement in American social science towards greater empiricism and 'realism'.

At its inception, then, institutionalism could be seen as a very promising programme – modern, scientific, pointing to a critical investigation and analysis of the existing economic system and its performance, in tune with the latest in psychological, social scientific, and legal research, established at leading universities and research institutes, and involved in important issues of economic policy and reform (see also Yonay 1998).

The Contributions of Interwar Institutionalism

Mark Blaug has stated that institutionalism 'was never more than a tenuous inclination to dissent from orthodox economics' (Blaug 1978, p. 712), and this view still finds wide currency. In fact, institutionalsm in the inter-war period was a major part of a pluralistic mainstream economics (Morgan and Rutherford 1998). That institutionalists did have a positive programme of research in mind should be clear from the above. Not all elements of this programme were pursued successfully, but there can be no doubt that institutionalists did make important positive contributions to economics, and this is particularly true of the period when institutionalism was at its peak. Just a few of these contributions will be highlighted.

Institutionalist took the task of improving economic measurement seriously. The NBER not only produced many empirical studies relating to business cycles, labour, and price movements, but also played a vital role in the development of national income accounting, through the work of Mitchell's student, Simon Kuznets. In conjunction with the Federal Reserve, the NBER also did much to develop monetary and financial data. Moreover, during the New Deal, institutionalists were heavily involved in the effort to improve the statistical work of government agencies (Rutherford 2002).

As noted above, one of the claims of institutionalists was that a 'scientific' economics would have to be consistent with 'modern' psychology. A typical argument was that economics 'is a science of human behaviour' and any conception of human behaviour that the economist may adopt 'is a matter of psychology' (Clark 1918, p. 4). Clark made one of the most interesting efforts to develop the psychological basis of institutional economics. Building on the work of William James and Cooley, he argued that the 'effort of decision' is an important cost, and one that prevents maximization. Clark was considering both the costs of information gathering and of calculation, and his argument is a clear precursor of more recent conceptions of bounded rationality leading to the use of habits or routines.

Interesting work on the economics of consumption and the household, was pursued by Hazel Kyrk and Theresa McMahon. McMahon made use of Veblen's conception of emulation in consumption, while Kyrk was critical of marginal utility theory as a basis for a theory of consumption and emphasized the social nature of the formation of consumption values. Consumption patterns relate to habitual 'standards of living', and Kyrk undertook to measure and critically analyse existing standards of living, and to create policy to help achieve higher standards of living. In her later work she discussed the household in both its producing and consuming roles, the division of labour between the sexes, employment and earnings of women, adequacy of family incomes, and issues of risks of disability, unemployment, provision for the future and social security, and the protection and education of the consumer (Kyrk 1923; 1933; McMahon 1925).

There was much work dealing with the inadequacy of the standard models of perfect competition and pure monopoly. The soft coal industry received particular attention. In that industry investigators such as Hamilton found little that corresponded to the ideal of a competitive industry. Competition within the industry had resulted not in efficient low-cost production but in persistent excess capacity, inefficiency, irregular operation, poor working conditions and low earnings (Hamilton and Wright 1925). This represented a common institutionalist theme – that, particularly under conditions of high overheads and rapid technological advance, competition could lead to 'disorder' and inefficiency rather than to order and efficiency. Institutionalists also studied such things as common pool problems in the oil industry, production cycles in agriculture, including the cobweb model and its implications for the orthodox view of 'self-regulating' markets, and the vast array of restrictive practices to be found in many industries (Hamilton and Associates 1938).

A related theme was that technological change had altered the structure of costs faced by firms and had altered their behaviour. This argument derived from Clark's *Overhead Costs* (1923). For Clark, the growth of overhead costs as a result of capital-intensive methods of production had resulted in price discrimination, an extension of monopoly and an increase in price inflexibility over the cycle. A little later Gardiner Means (1935) developed his theory of administered pricing, which sparked a vast literature on relative price inflexibility.

On issues of corporate finance and ownership, Bonbright and Means co-authored *The Holding Company*, and Berle and Means *The Modern Corporation and Private Property*, both in 1932. These works much extended Veblen's earlier discussions of corporate consolidation and the separation of ownership and control. Berle and Means's work raised important issues of agency, and whether managers would maximize profits.

On labour market issues, institutionalists concerned themselves with studying unions and the history of the labour movement, developing in

the process both classifications of unions and explanations for the particular pattern of trade union development in America (Perlman 1928). Wage determination was also a problem that attracted the attention of institutionalists. Walton Hamilton's 1923 book *The Control of Wages* (with Stacy May) was praised by Clark for providing not an 'abstract formulation of the characteristic outcome' but a 'directory of the forces to be studied' in any particular case (Clark 1927, pp. 276–7). Discussions of trade unions and wage bargaining were provided by other institutional labour economists such as Commons (1924) and Sumner Slichter (1931). In this work much attention was given to issues of collective bargaining and systems of conciliation and mediation.

Public utilities, including issues relating to the valuation of utility property and the proper basis for rate regulation, were major areas of institutionalist research. Both Clark and Commons devoted considerable attention to the concept of intangible property, goodwill, and valuation issues (Commons 1924; Clark 1926). Bonbright dealt with the difference between commercial and social valuation in connection with public utilities. Bonbright, Hale, and Martin Glaeser all wrote extensively on issues of public utility regulation, with Hale probably having the greatest impact with his campaign of criticism of the 'fair value' concept as a basis for rate regulation (Hale 1921; Bonbright 1961, p. 164).

In his *Social Control of Business* (1926) Clark argued that business cannot be regarded as a purely private affair. This idea of private business being broadly 'affected with a public interest' was absolutely central to the institutionalist argument for regulation of business. Clark expresses the idea in his claim that 'every business is "affected with a public interest" of one sort or another' (Clark 1926, p. 185), and the argument also appears in as a central theme in Tugwell's early work on regulation (Tugwell 1921, 1922), and in Walton Hamilton's and Robert Hale's extensive writings on law and economics (Rutherford 2005b; Fried 1998).

More general interconnections between law and economics and the operation of markets were addressed by Hale, Commons, and Hamilton. Commons's approach was the most developed and was built on his notions of the pervasiveness of distributional conflicts, of legislatures and courts as attempting to resolve conflicts (at least between those interest groups with representation), and of the evolution of the law as the outcome of these ongoing processes of conflict resolution. He developed his concept of the 'transaction' as the basic unit of analysis (later adopted by Oliver Williamson). In turn, the terms of transactions were determined by legal rights and by economic (bargaining) power. Market transactions always involved some degree of 'coercion', in the sense of some degree of restriction upon alternatives (Commons 1924, 1932; Hale 1923). He also provided a theory of the behaviour of legislatures based on 'log-rolling', and a theory of judicial decision-making based on the concept of 'reasonableness', a concept that included, but was not limited to, a concern with efficiency (Commons 1932; 1934).

The institutionalist programme dealing with business cycles, in the period before the depression, was centred on Wesley Mitchell's work and that he promoted through the NBER. As noted above, Mitchell explicitly placed his work on business cycles within an institutional context by associating cycles with the functioning of the system of pecuniary institutions. Mitchell's 1913 volume *Business Cycles*, with its discussion of the four-phase cycle driven by an interaction of factors such as the behaviour of profit seeking firms, the behaviour of banks, and the leads and lags in the adjustment of prices and wages, became the standard institutionalist reference. At the NBER, Mitchell focused heavily on promoting work that would add to the understanding of business cycles, generating a stream of research studies far too long to list here, but contributing to the development of national income measures, business cycle indicators, and much more. In addition, Clark developed his concept of the accelerator out of his study of Mitchell's 1913 work, and the accelerator mechanism soon became a standard part of cycle theory (Clark 1917). Mitchell's work was not the only approach to business cycles to be found within institutionalism. Many

institutionalists, including Hamilton, had an interest in the work of J.A. Hobson, and Hobson's underconsumptionism became popular among institutionalists in the 1930s (Rutherford 1994).

On issues of market failure, broadly conceived, Clark (1926) discussed a large number of types of market failure in his *Social Control of Business*. These included monopoly, maintaining the ethical level of competition, protecting individuals where they are unable to properly judge alternatives, problems of agency, relief for people displaced by rapid economic and technological change, relief of poverty (including social security and minimum wages), regulation of advertising and the provision of information and standards, increasing equality of opportunity, externalities ('unpaid costs of industry'), public goods ('inappropriable services'), the wastes of 'arms race' types of competition (such as competitive advertising), unemployment, the interests of posterity or future generations, and any other discrepancy between private and social accounting. Slichter (1924) provided a list of problems almost as long, including the pro-cyclical behaviour of banks, overexploitation of natural resources, discrimination in employment, advertising and salesmanship, lack of market information, pollution and other external effects, uncertainty and unemployment, economic waste and inefficiency, and economic conflict. All these problems were seen as justifying some additional 'social control' of business activity.

Finally, and intimately related to the above, institutionalists made important contributions to policy in their roles in the development of unemployment insurance, workmen's compensation, social security, labour legislation, public utility regulation, agricultural price support programmes, and in the promotion of government 'planning' to create high and stable levels of output. Commons had pioneered public utility regulation, unemployment insurance, and workmen's compensation in Wisconsin, and the Wisconsin model was widely influential. Many institutionalists were active members of the American Association of Labor Legislation (AALL), and the AALL promoted many reforms to labour legislation. Medical insurance programmes were also pursued by the AALL, and also by the Committee on the Cost of Medical Care, which involved both Hamilton and Mitchell.

Institutionalists had significant influence within the New Deal. Many of Commons's students played leading roles in the development of the federal social security programme. Berle and Tugwell were two of Roosevelt's original 'Brains Trust', and Tugwell, Means, and Mordecai Ezekiel were the leading advocates of the 'structuralist' or planning approach that had influence in the early part of the New Deal (Barber 1996). Hamilton and several others were deeply involved in the labour legislation and consumer protection aspects of the New Deal. Hamilton later worked with Thurman Arnold in developing their case by case approach to anti-trust (Rutherford 2005b).

Institutional Economics After 1945

Institutionalism attained a significant position in American economics in the interwar period, both in academia and in government, but then declined in position and prestige after the Second World War. At this point institutionalism fell out of the mainstream of American economics to become a heterodox tradition on the margins of the discipline. There are quite a number of overlapping reasons for this, some of which reach back into the 1920s and 1930s, but the focus here will be limited to just a few of the more important issues.

Institutionalism clearly did not live up to its own early promise, particularly in its failure to pin down exactly what foundations in 'modern psychology' it was supposed to have. After the mid-1920s, psychologists abandoned the instinct/habit approach in favour of a behaviourism that became increasingly narrow and difficult to see as an adequate foundation for economics. In this climate, the enthusiasm for new psychological approaches that had played such a role in the institutionalist movement's beginnings could not be sustained. Institutionalism probably played a part in ridding economics of explicitly hedonistic language, but it did not develop the alternative

basis to convince the profession as a whole to abandon its traditional views of rationality (Lewin 1996).

It must also be said that institutionalists failed to develop their theories of social norms, technological change, legislative and judicial decision-making, transactions, and forms of business enterprise (apart from issues of ownership and control) much beyond the stage reached by Veblen and Commons. The reasons for this lack of development relate partly to the focus of interwar institutionalists on immediate and pressing policy problems, like business cycles, labour law, and social security. In addition, from the late 1920s on, sociology separated itself from economics and became established in separate departments, taking much of the subject matter of social norms and institutions with it.

It is also the case that, from the 1930s onwards, many new developments in theory and methods occurred within economics: developments that tended to displace institutionalist ideas and methods. Hicks's revision of demand theory seemed to free economics from the shifting basis of psychology, while the work of Joan Robinson and Edward Chamberlin provided treatments of imperfect competition more amenable to neoclassical approaches. The discussion of externalities in terms of market failure was also much clarified. Neoclassicism developed a language capable of encompassing at least some of the issues of concern to institutionalists; issues that had formerly fallen outside the neoclassical theoretical compass.

Moreover, institutionalist approaches to business cycles were replaced by Keynesian ideas. In many respects, Keynesian economics took over the role of the exciting 'new' economics that institutionalism had played in the early 1920s. In addition, neoclassical and Keynesian economics gained an empirical component with the rise of econometrics. Institutionalists could no longer claim greater 'scientific' standing because of their empiricism; indeed, they were accused by Koopmans (1947) of 'measurement without theory'; a much exaggerated view, but one often repeated and widely accepted.

In these ways more 'orthodox' economic theory took over those aspects of institutionalism amenable to 'model analysis' (Copeland 1951) while other aspects were absorbed into what became applied field areas, such as industrial organization, labour economics, and industrial relations. At least until the 1960s these field areas had only loose ties to the theoretical core of the discipline, and maintained a substantial institutional component.

Finally, a significant part of the institutionalist agenda of social reform had come to pass, both removing some of the original causes of the institutionalist movement, and prompting a reaction in the form of critiques of the expanded role for government that institutionalists had done so much to put forward.

Under these circumstances, it is not difficult to see why institutionalism slipped from being a central part of American economics to a more marginalized position. This change did not happen overnight, but was hastened by the significant amount of new hiring on the part of American universities immediately after the Second World War. These new faculty were predominantly Keynesians or neoclassicals equipped with the latest in mathematical and econometric tools. The retirement of the last of the older generation of institutionalists in the 1950s completed the process.

American institutionalism did not disappear, but it certainly changed. Insitutionalists formed the small 'Wardman Group' in 1959, an organization that later became the Association for Evolutionary Economics, still the primary organization of 'old' institutionalists in America, and the publisher of *The Journal of Economic Issues*. Institutionalism disassociated itself from the positivism that had gained popularity elsewhere (a positivism that, ironically, Mitchell and the NBER had played an important part in creating), and turned away from the methods and the core areas of the discipline that had been taken over by neoclassical and Keynesian economics. Institutionalists continued to work in applied areas, and to argue for more active government regulation and 'planning' of the economy (Gruchy 1974), but there was also something of a movement back to the broader institutional themes found in Veblen and Commons.

This tendency was especially promoted by Clarence Ayres, in his *Theory of Economic Progress* (1944). Ayes attempted to renew the Veblenian emphasis on technology as the driving force behind institutional change, and developed the Veblenian distinction between business and industry into a general dichotomy between the ceremonial and instrumental aspects of culture. Ayres's charismatic personality attracted a number of students to the institutionalist ranks, and they spread his version of institutionalism to many south-western universities. The University of Texas, too, retained its institutionalist character longer than most, and in the 1960s was still the home of a substantial institutionalist group. Other institutionalist groups existed at Maryland and at Michigan State. J.K. Galbraith produced widely read and distinctly Veblenian analyses in his *Affluent Society* (1958) and *New Industrial State* (1971), while the Commons tradition in law and economics has been kept alive by Daniel Bromley, Allan Schmid, and Warren Samuels (Samuels 1971; Schmid 1978; Bromley 1989).

Perhaps the most important recent development within the 'old' institutionalist tradition has been the growing interest in the work of Veblen and Commons among a new generation of European economists attracted to institutional and evolutionary ideas. One outstanding example of this is to be found in the work of Geoffrey Hodgson, who has argued forcefully for the development of an institutional economics along lines he sees as having been originally pioneered by Veblen in his evolutionary and Darwinian approach to institutions and institutional change (Hodgson 1988; 2004).

See Also

- ▶ Ayres, Clarence Edwin (1891–1972)
- ▶ Clark, John Maurice (1884–1963)
- ▶ Commons, John Rogers (1862–1945)
- ▶ Mitchell, Wesley Clair (1874–1948)
- ▶ United States, economics in (1885–1945)
- ▶ Veblen, Thorstein Bunde (1857–1929)

Bibliography

Ayres, C.E. 1944. *The theory of economic progress*, 2nd ed, 1962. New York: Schocken.

Barber, W.J. 1996. *Designs within disorder: Franklin D. Roosevelt, the economists, and the shaping of economic policy, 1933–1945*. New York: Cambridge University Press.

Blaug, M. 1978. *Economic theory in retrospect*, 3rd ed. London: Cambridge University Press.

Bonbright, J.C. 1961. *Principles of public utility rates*. New York: Columbia University Press.

Bromley, D.W. 1989. *Economic interests and institutions: The conceptual foundations of public policy*. Oxford: Basil Blackwell.

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

Clark, J.M. 1918. Economics and modern psychology, I and II. *Journal of Political Economy* 26(1–30): 136–166.

Clark, J.M. 1919. Economic theory in an era of social readjustment. *American Economic Review* 9: 280–290.

Clark, J.M. 1921. Soundings in non-Euclidean economics. *American Economic Review* 11: 132–143.

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

Clark, J.M. 1926. *Social control of business*. Chicago: University of Chicago Press.

Clark, J.M. 1927. Recent developments in economics. In *Recent developments in the social sciences*, ed. E.C. Hayes. Philadelphia: Lippencott.

Commons, J.R. 1924. *The legal foundations of capitalism*. Madison: University of Wisconsin Press, 1968.

Commons, J.R. 1932. The problem of correlating law, economics and ethics. *Wisconsin Law Review* 8: 3–26.

Commons, J.R. 1934. *Institutional economics: Its place in political economy*. New York: Macmillan.

Copeland, M.A. 1951. Institutional economics and model analysis. *American Economic Review* 41: 56–65.

Edie, L.D. 1927. Some positive contributions of the institutional concept. *Quarterly Journal of Economics* 41: 405–440.

Fried, B.H. 1998. *The progressive assault on Laissez Faire: Robert Hale and the first law and economics movement*. Cambridge, MA: Harvard University Press.

Galbraith, J.K. 1958. *The affluent society*. Boston: Houghton Mifflin.

Galbraith, J.K. 1971. *The new industrial state*, 2nd ed. Boston: Houghton Mifflin.

Gruchy, A.G. 1974. Government intervention and the social control of business: The neoinstitutionalist position. *Journal of Economic Issues* 8: 235–249.

Hale, R.L. 1921. The "physical value" fallacy in rate cases. *Yale Law Journal* 30: 710–731.

Hale, R.L. 1923. Coercion and distribution in a supposedly non-coercive state. *Political Science Quarterly* 38: 470–494.

Hamilton, W.H. 1919. The institutional approach to economic theory. *American Economic Review* 9: 309–318.
Hamilton, W.H., and S. May. 1923. *The control of wages*, 1968. New York: Augustus M. Kelley.
Hamilton, W.H., and H.R. Wright. 1925. *The case of bituminous coal*. New York: Macmillan.
Hamilton, W.H. and Associates. 1938. *Price and price policies*. New York: McGraw Hill.
Hodgson, G.M. 1988. *Institutions and economics: A manifesto for a modern institutional economics*. Cambridge: Polity Press.
Hodgson, G.M. 2004. *The evolution of institutional economics: Agency, structure and Darwinism in American institutionalism*. London: Routledge.
Koopmans, T.C. 1947. Measurement without theory. *Review of Economic Statistics* 29: 161–172.
Kyrk, H. 1923. *A theory of consumption*. Boston: Houghton Mifflin.
Kyrk, H. 1933. *Economic problems of the family*. New York: Harper.
Lewin, S. 1996. Economics and psychology: Lessons for our own day from the early twentieth century. *Journal of Economic Literature* 35: 1293–1323.
McMahon, T. 1925. *Social and economic standards of living*. Boston: D.C. Heath.
Means, G.C. 1935. *Industrial prices and their relative inflexibility*. Senate Document 13. In: 74th congress, 1st session. Washington, DC: US Government Printing Office.
Mitchell, W.C. 1910a. The rationality of economic activity, I. *Journal of Political Economy* 18: 97–113.
Mitchell, W.C. 1910b. The rationality of economic activity, II. *Journal of Political Economy* 18: 197–216.
Mitchell, W.C. 1913. *Business cycles*. Berkeley: University of California Press.
Mitchell, W.C. 1924. The prospects of economics. In *The trend of economics*, ed. R.G. Tugwell. Port Washington: Kennikat Press, 1971.
Mitchell, W.C. 1925. Quantitative analysis in economic theory. *American Economic Review* 15: 1–12.
Mitchell, W.C. 1927. *Business cycles: The problem and its setting*. New York: NBER.
Morgan, M.S. and Rutherford, M. 1998. *From interwar pluralism to postwar neoclassicism*. Annual Supplement to vol. 30 of History of Political Economy, Durham, NC: Duke University Press.
Perlman, S. 1928. *A theory of the labor movement*. New York: Macmillan.
Rutherford, M. 1994. J.A. Hobson and American institutionalism: Underconsumption and technological change. In *J.A. Hobson after fifty years*, ed. J. Pheby. London: Macmillan.
Rutherford, M. 1999. Institutionalism as 'scientific' economics. In *From classical economics to the theory of the firm: Essays in honour of D.P. O'Brien*, ed. R. Backhouse and J. Creedy. Aldershot: Edward Elgar.
Rutherford, M. 2000a. Institutionalism between the wars. *Journal of Economic Issues* 34: 291–303.
Rutherford, M. 2000b. Understanding institutional economics: 1918–1929. *Journal of the History of Economic Thought* 22: 277–308.
Rutherford, M. 2001. Institutional economics: Then and now. *Journal of Economic Perspectives* 15: 173–194.
Rutherford, M. 2002. Morris A. Copeland: A case study in the history of institutional economics. *Journal of the History of Economic Thought* 24: 261–290.
Rutherford, M. 2003. Walton Hamilton, Amherst, and the Brookings graduate school: Institutional economics and education. *History of Political Economy* 35: 611–653.
Rutherford, M. 2004. Institutional economics at Columbia University. *History of Political Economy* 36: 31–78.
Rutherford, M. 2005a. Who's afraid of Arthur Burns? The NBER and the foundations. *Journal of the History of Economic Thought* 27:109–139
Rutherford, M. 2005b. Walton H. Hamilton and the public control of business. In *The role of government in the history of political economy*, ed. S. Medema and P. Boettke. Supplement to volume 37, *History of Political Economy*. Durham: Duke University Press.
Rutherford, M. 2006. Wisconsin institutionalism: John R. Commons and his students. *Labor History* 47: 161–188.
Rutherford, M. 2007. Chicago economics and institutionalism. In *Elgar companion to Chicago economics*, ed. R. Emmett. Aldershot: Edward Elgar (forthcoming).
Samuels, W.J. 1971. The interrelations between legal and economic processes. *Journal of Law and Economics* 14: 435–450.
Schmid, A.A. 1978. *Property, power, and public choice*. New York: Praeger.
Slichter, S.H. 1924. The organization and control of economic activity. In *The trend of economics*, ed. R.G. Tugwell. Port Washington: Kennikat Press, 1971.
Slichter, S.H. 1931. *Modern economic society*. New York: H. Holt.
Stewart, W.W. 1919. Economic theory: Discussion. *American Economic Review* 9: 319–320.
Tugwell, R.G. 1921. The economic basis for business regulation. *American Economic Review* 11: 643–658.
Tugwell, R.G. 1922. *The economic basis of public interest*. New York: Augustus M. Kelley, 1968.
Tugwell, R.G. 1924. *The trend of economics*. Port Washington: Kennikat Press, 1971.
Veblen, T. 1898. Why is economics not an evolutionary science? In *The place of science in modern civilisation*. New York: Russell & Russell, 1961.
Veblen, T. 1899. *The theory of the leisure class*, 1924. London: George Allen & Unwin.
Veblen, T. 1904. *The theory of business enterprise*. Clifton: Augustus M. Kelley, 1975.
Yonay, Y.P. 1998. *The struggle over the soul of economics: Institutionalist and neoclassical economists in America between the wars*. Princeton: Princeton University Press.

Instrumental Variables

Charles E. Bates

Abstract

Instrumental variables methods are an essential tool in modern econometric practice. The method itself is of ancient lineage and historically is closely connected with the econometrics of simultaneous equations. This article describes the statistical foundations of instrumental variables methods with a focus on their classical development.

Keywords

Central limit theorems; Errors in variables; Euler equations; Generalized method of moments estimation; Instrumental variables; Law of large numbers; Natural experiments; Returns to schooling; Serial correlation; Simultaneous equations models; Treatment effect; Two-stage least squares estimator

JEL Classifications

C

In one of its simplest formulations the problem of estimating the parameters of a system of simultaneous equations with unknown random errors reduces to finding a way of estimating the parameters of a single linear equation of the form $Y = X\beta_0 + \varepsilon$, where β_0 is unknown, Y and X are vectors of data on relevant economic variables and ε is the vector of unknown random errors. The most common method of estimating β_0 is the method of least squares:

$\hat{\beta}_{\mathrm{OLS}} \equiv \arg\min \varepsilon(\beta)'\varepsilon(\beta)$, where $\varepsilon(\beta) \equiv Y - X\beta$. Under fairly general assumptions $\hat{\beta}_{\mathrm{OLS}}$ is an unbiased estimator of β_0 provided $E(\varepsilon_t|X) = 0$ for all t, where ε_t is the tth-coordinate of ε.

Unfortunately for the empirical economist, it is often the case that the basic orthogonality condition between the errors and the explanatory variables is not satisfied by economic models, due to correlation between the errors and the explanatory variables. Particularly relevant examples of this situation include (1) any case where the data contain errors introduced by the process of collection (errors in variables problem); (2) the inclusion of a dependent variable of one equation in a system of simultaneous equations as an explanatory variable in another equation in the system (simultaneous equations bias); and (3) the inclusion of a lagged dependent variable as an explanatory variable in the presence of serial correlation. For all of these cases,

$$E\left(\widehat{\beta}_{\mathrm{OLS}}\right) = E\left[(X'X)^{-1}X'Y\right]$$
$$= \beta_0 + E\left[(X'X)^{-1}\sum_{t=1}^{n} X_t' E(\varepsilon_t|X)\right] \neq \beta_0$$

in general, and the bias introduced cannot be determined because the errors ε are unknown. Furthermore, in every case the bias fails to go to zero as the sample size increases. Clearly the method of least squares is unsatisfactory for many situations of relevance to economists.

In 1925 the US Department of Agricultural published a study by the zoologist Sewall Wright where the parameters of a system of 6 equations in 13 unknown variables were estimated using a method he referred to as 'path analysis'. In essence his approach exploited zero correlations between variables within his system of equations to construct a sufficient number of equations to estimate the unknown parameters. The idea which underlies this approach is that, if two variables are uncorrelated, then the average of the product of repeated observations of these variables will approach zero as the number of observations is increased without bound except for a negligible number of times. Thus if we know that a variable of the system Z_i is uncorrelated with the errors ε, we can exploit the fact that $n^{-1}\sum_{t=1}^{n} Z_{ti}\varepsilon_t \equiv n^{-1}\sum_{t=1}^{n} Z_{ti}(Y_t - X_t\beta_0)$ approaches zero to construct a useful relationship between parameters of the system by setting such averages equal to zero. Provided a sufficient number of such relationships can be constructed which are independent, this

provides a method for estimating the parameters of a system of simultaneous equations which should become more accurate as the number of observations increases.

Since the 1940s, when Reiersøl (1941, 1945) and Geary (1949) presented the formal development of this procedure, the variables Z which are instrumental in the estimation of the parameters β_0 have been called 'instrumental variables'. Associated with each instrumental variable Z_i is an equation formed as described in the previous paragraph, called a normal equation, which can be used to form the estimates of the unknown parameters. Frequently there are more instrumental variables than parameters to be estimated. As the equations are formed from relationships between random variables, generally no solution will exist to a system of estimating equations formed in this manner using all possible instrumental variables. As each estimating equation contains relevant information about the parameters to be estimated, it is undesirable just to ignore some of them. Thus we can define a fundamental problem in the application of this method: how can we make effective use of all the information available from the instrumental variables? This problem will occupy the rest of this article.

Let $\varepsilon_t(\theta) \equiv F_t(X_t, Y_t, \theta)$ be a $p \times 1$ vector-valued function defined on a domain of possible parameter values $\Theta \subseteq \mathbb{R}^k$ which represents a system of p simultaneous equations with dependent variables Y_t, a $p \times 1$ random vector, and an $m \times s$ random matrix of explanatory variables X_t for all $t = 1, 2, ..., n$. Standard formulations of $F_t(X_t, Y_t, \theta)$ are the linear model $\varepsilon_t(\theta) \equiv Y_t - X_t\theta$ and the nonlinear model $\varepsilon(\theta) \equiv Y_t - f_t(X_t, \theta)$. Let $W_t(\theta)$ be a $p \times r$ random valued matrix defined on Θ for all $t = 1, 2,..., n$. Assume that there exists a unique value θ_0 in Θ such that $E\left(\varepsilon_1^0 | W_1^0\right) = 0$ for all $t = 1, 2, \ldots, n$, where $\varepsilon_t^0 \equiv F_t(X_t, Y_t, \theta_0)$ and $W_t^0 \equiv W_t(\theta_0)$. Finally, let $Z_t(\theta)$ be a $p \times 1$ random matrix such that $E(Z_t(\theta)|W_t(\theta) = Z_t(\theta)$ for all θ in Θ. Any such variables $Z_t(\theta_0)$ may serve as instrumental variables for the estimation of the unknown parameters θ_0 since

$$E\left(Z_t^{0\prime}\varepsilon_t^0\right) = E\left(E\left(Z_t^{0\prime}\varepsilon_t | W_t^0\right)\right) = E\left(Z_t^0 E\left(\varepsilon_t^0 | W_t^0\right)\right) = 0$$

for all $t = 1, 2, \ldots, n$, as long as the functions F_t and W_t and the data generating process satisfy sufficiently strong regularity assumptions to ensure that the uniform law of large numbers is satisfied, that is

$$n^{-1}\sum_{t=1}^{n} Z_t(\theta)'\varepsilon_t(\theta) \to n^{-1}\sum_{t=1}^{n} E\left[Z_t(\theta)'\varepsilon(\theta)\right]$$

uniformly in θ on Θ.

Identification of the unknown parameters θ_0 requires that there be at least as many instrumental variables as there are parameters to be estimated, that is, $l \geq k$. On the other hand, if there are more instrumental variables than parameters to be estimated, there will be no solution to $n^{-1}\sum_{t=1}^{n} Z_t(\theta)'\varepsilon_t(\theta) = 0$ in general for finite n as indicated above. One possible solution to this problem is simply to use k of the instrumental variables in the estimation of θ_0. The omitted instrumental variables may then be used to construct statistical tests of the $l - k$ overidentifying restrictions of the unknown parameter vector. A drawback of this approach is that not all of the information available to us is used in the estimation of the unknown parameters and hence, the estimates will not be as precise as they should be. An alternative approach which effectively uses all of the available instrumental variables is to be preferred.

Even though in general the moment function $n^{-1}\sum_{t=1}^{n} Z_t'(\theta)\varepsilon_t(\theta) \neq 0$ for any value of θ, its limiting function $n^{-1}\sum_{t=1}^{n} E\left[Z_t'(\theta)\varepsilon_t(\theta)\right]$ does vanish when $\theta = \theta_0$. This suggests estimating θ_0 with that of Θ which makes $n^{-1}\sum_{t=1}^{n} Z_t'(\theta)\varepsilon_t(\theta)$ as close to zero as possible. The criterion of closeness is of some interest to the econometrician. It affects the size of the confidence ellipsoids of the estimator about θ_0 and hence the precision of the estimate. The nonlinear instrumental variables estimator (NLIV), $\hat{\theta}_{n,\mathrm{NLIV}} = \mathrm{argmin}_{\theta \in \Theta}$

$$\left[\sum_{t=1}^{n} Z_t'(\theta)\epsilon_t(\theta)\right]' \cdot \left[\mathrm{Var}\sum_{t=1}^{n} Z_t'(\theta_0)\epsilon_t(\theta_0)\right]^{-1} \cdot \left[\sum_{t=1}^{n} Z_t'(\theta)\epsilon_t(\theta)\right]$$

is the optimal instrumental variables estimator in this respect (Bates and White 1986a).

The NLIV estimator simplifies to well-known econometric estimators in a variety of alternative specifications of the underlying probability model which generated the variables. When the data generating process is independent and identically distributed, $\hat{\theta}_{n,\mathrm{NLIV}}$ is the nonlinear three-stage least squares estimator of Jorgenson and Laffont (1974). The additional restriction of consideration to a single equation ($p = 1$) results in the nonlinear two-stage least squares estimator of Amemiya (1974). Furthermore, if the model $\varepsilon(\theta)$ is linear in $\theta, \hat{\theta}_{n,\mathrm{NLIV}}$ then simplifies to the three-stage least squares estimator of Zellner and Theil (1962) for a system of simultaneous equations and to the two-stage least squares estimator of Theil (1953), Basmann (1957) and Sargan (1958) for the estimation of the parameters of a single equation. On the other hand, if we allow for heterogeneity by restricting the data generating process only to be independent, $\hat{\theta}_{n,\mathrm{NLIV}}$ simplifies to White's (1982) two-stage instrumental variables estimator of the parameters of a single linear equation.

As indicated above, it is desirable from consideration of asymptotic precision to include as many instrumental variables as are available for the estimation of the unknown parameters θ_0. This raises the question of the existence of a set of instrumental variables $\{Z^*\} \in \Gamma$ that renders the inclusion of any further instrumental variables redundant, where Γ is the set of all sequences of instrumental variables such that $\hat{\theta}_{n,\mathrm{NLIV}}$ is a consistent estimator of θ_0 with an asymptotic covariance matrix. Bates and White (1986b) provide conditions which imply that such instrumental variables exist, though it may not be possible to obtain them in practice. Suppose there exists a sequence of k instrumental variables $\{\mathcal{L}\}$ such that for all $\{Z\}$ in Γ

$$E[\mathcal{L}(\theta_0)'\nabla_\theta \varepsilon(\theta_0)] = E\left[\mathcal{L}(\theta_0)'\varepsilon(\theta_0)\varepsilon(\theta_0)'Z(\theta_0)\right].$$

Then $\mathcal{L}(\theta_0)$ is optimal in Γ in the sense of asymptotic precision. Suppose it is also the case that Σ is an $np \times np$ matrix with representative element $\sigma_{th\tau g} = E(\varepsilon_{\mathrm{th}}(\theta_0) \cdot \varepsilon_{\tau g}(\theta_0) \quad | W_{th}(\theta_0), W_{\tau g}(\theta_0))$, is nonsingular a.s. and that

$$E\left[E(\nabla_\theta \varepsilon_{th}(\theta_0)|\, W_{th}(\theta_0))|\, W_{\tau g}\right] = E(\nabla_\theta \varepsilon_{th}(\theta_0)|\, W_{th}(\theta_0))$$

for all t, τ=1,2,..., n and h, g = 1,2,..,p such that $\sigma^{th\tau g} \equiv 0$, where $\sigma^{th\tau g}$ is a representative element of Σ^{-1}. Let Z^* be an $np \times k$ matrix with rows

$$Z^*_{th} \equiv \sum_{\tau=1}^{n}\sum_{g=1}^{p} \sigma^{th\tau g} E\left[\nabla_\theta \varepsilon_{\tau g}(\theta_0)|\, W_{\tau g}(\theta_0)\right].$$

If $\{Z^*\}$ is in Γ then $\{Z^*\}$ is optimal in Γ.

In many situations it will not be possible to make use of such instrumental variables in practice. However, for some important situations optimal instrumental variables are available. Suppose that $\varepsilon(\theta) \equiv Y - X\theta$ and the explanatory variables X are independent of the errors $\varepsilon(\theta_0)$. If the errors are independent and identically distributed for all $t = 1, 2,...,n$ and $h = 1, 2, \ldots, p$, then $Z^* = X$. Thus the optimal instrumental variables estimator is given by

$$\hat{\theta}_{n,\mathrm{NLIV}} \equiv \arg\min_{\theta \in \Theta} \varepsilon(\theta)'X\left[\sigma^2 E(X'X)\right]^{-1}X'\varepsilon(\theta),$$

where $\sigma^2 \equiv \mathrm{var}[\varepsilon_{\mathrm{th}}(\theta_0)]$ is a real, nonstochastic scalar for all t and h. If it is also the case that $n^{-1}E(X'X) - n^{-1}X'X \to 0$ as $n \to \infty$, $\hat{\theta}_{n,\mathrm{NLIV}}$ is asymptotically equivalent to $\arg\min_{\theta \in \Theta}\varepsilon'(\theta)X(X'X)^{-1}X'\varepsilon(\theta)$, that is ordinary least squares is the optimal instrumental variables estimator. If there is contemporaneous correlation only, that is, $\mathrm{var}(\varepsilon_t(\theta_0)) = \Omega$, a $p \times p$ nonstochastic matrix, then Zellner's (1962) seemingly unrelated regression estimator (SURE), is the optimal instrumental variables estimator. If we further relax these assumptions so that $\mathrm{var}(\varepsilon(\theta_0))$ is an arbitrary positive definite $np \times np$ matrix, the generalized least squares (Aitken 1935) is the optimal instrumental variables estimator.

Since the development of the two-stage least squares estimator in the mid-1950s, the method of instrumental variables has come to play a prominent role in the estimation of economic relationships. In turn, in modern econometric practice, the use of instrumental variables methods has been very much influenced by the evolution of

economic theory as well as the evolution of the relationship between theory and empirical practice. Within macroeconomics, the standard approach to structural economic analysis, Hansen's (1982) generalized method of moments estimation (GMM) procedure, which is a generalization of nonlinear instrumental variables estimation, typically relies on economic theory to identify validinstruments. The most prominent application of GMM methods is to Euler equations estimation, in which valid instruments are defined by a theoretical specification of how one set of variables may be understood to be the expected value of another set.

While economic theory can identify instruments that are, in principle, valid, it typically does not provide insights into whether such instruments are strongly or weakly correlated with the variables that are to be instrumented. Concern over the problem of weak instruments was stimulated by the demonstration in Bound, Jaeger and Baker (1995) that the failure to account for weak instruments called into question findings on the return to schooling in the work of Angrist and Krueger (1991). Important analyses of the effects of weak instruments on estimation and inference include Dufour and Taamouti (2005), Staiger and Stock (1997) and Stock and Wright (2000). This new literature is well surveyed in Dufour (2003) and Andrews and Stock (2006). This work demonstrates how lack of attention to the possibility of weak instruments can lead to very misleading inferences for a broad range of contexts. Furthermore, it provides new ways for conducting inference with respect to the parameters of interest.

When economic theory identifies valid instruments, it may be the case that the set of potential instruments is unbounded. This is evident in Euler equation contexts, where the validity of the instrument vector Z_{t-k} usually implies the validity of Z_{t-k-l}, $l > 0$. This has led to research on the properties of estimators where the number of available instruments grows with the number of observations. Donald and Newey (2001) and Hahn (2002) are examples of analyses of this type.

Other developments in the use of instrumental variables have occurred because of a desire to avoid the use of structural models in developing substantive economic claims. In particular, the literature on TREATMENT EFFECTS may be understood as an effort to show how the causal effects of alternative policies may be uncovered outside the context of a structural model. A part of this literature has focused on the analysis of data from natural experiments and quasi-natural experiments, but, as argued by Heckman (1996), such analyses are in fact forms of instrumental variable estimation. Perhaps unsurprisingly, the effort to develop causal claims based on statistical rather than economic assumptions has led to controversy since the two types of assumptions are in fact interrelated. See Angrist et al. (1996) as well as the discussion of this paper for different perspectives. Heckman (1997) provides a wide-ranging critique of the use of instrumental variables in contemporary research; Roy model describes constructive ways to proceed.

Finally, it is important to remember that instrumental variable estimates are normally evaluated on the basis of asymptotic properties, that is, the law of large numbers and the central limit theorems. Since in general it is not possible to know how much data are required to arrive at acceptable estimates, conclusions derived from instrumental variables estimates should be tempered with a healthy dose of scepticism.

See Also

- ▶ Generalized Method of Moments Estimation
- ▶ Matching Estimators
- ▶ Natural Experiments and Quasi-natural Experiments
- ▶ Roy Model
- ▶ Treatment Effect
- ▶ Two-Stage Least Squares and the *k*-class Estimator

Bibliography

Aitken, A.C. 1935. On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh* 55: 42–48.

Amemiya, T. 1974. The nonlinear two-stage least-squares estimator. *Journal of Econometrics* 2: 105–110.

Andrews, D., and J. Stock. 2006. Inference with weak instruments. In *Advances in econometrics: Proceedings of the Ninth World Congress of the Econometric Society*, ed. R. Blundell, W. Newey, and T. Persson. Cambridge: Cambridge University Press.

Angrist, J., and A. Krueger. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106: 979–1014.

Angrist, J.D., G.W. Imbens, and D.B. Rubin. 1996. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 91: 444–472.

Basmann, R.L. 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25: 77–83.

Bates, C.E., and H. White. 1986a. Efficient estimation of parametric models. Working Paper No. 166, Department of Political Economy, Johns Hopkins University.

Bates, C.E., and H. White. 1986b. An asymptotic theory of estimation and inference for dynamic models. Working paper, Department of Political Economy, Johns Hopkins University.

Bound, J.D., D.A. Jaeger, and R. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90: 443–450.

Donald, S., and W. Newey. 2001. Choosing the number of instruments. *Econometrica* 69: 1365–1387.

Dufour, J.-M. 2003. Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics* 36: 767–808.

Dufour, J.-M., and M. Taamouti. 2005. Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica* 73: 1351–1365.

Geary, R.C. 1949. Determination of linear relations between systematic parts of variables with errors in observation, the variances of which are unknown. *Econometrica* 17: 30–58.

Goldberger, A.S. 1972. Structural equation methods in the social sciences. *Econometrica* 40: 979–1001.

Hahn, J. 2002. Optimal inference with many instruments. *Econometric Theory* 18: 140–168.

Hausman, J.A. 1983. Specification and estimation of simultaneous equation models. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 1. Amsterdam: North-Holland.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–1054.

Heckman, J. 1996. Randomization as an instrumental variable. *Review of Economics and Statistics* 78: 336–341.

Heckman, J. 1997. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* 32: 441–462.

Jorgenson, D.W., and J. Laffont. 1974. Efficient estimation of nonlinear simultaneous equations with additive disturbances. *Annals of Economic and Social Measurement* 3: 615–640.

Reiersøl, O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9: 1–24.

Reiersøl, O. 1945. Confluence analysis by means of instrumental sets of variables. *Arkiv for Mathematik, Astronomi och Fysik* 32A: 1–119.

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415.

Staiger, D., and J. Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65: 557–586.

Stock, J., and J. Wright. 2000. GMM with weak instruments. *Econometrica* 68: 1055–1096.

Theil, H. 1953. *Estimation and simultaneous correlation in complete equation systems*. The Hague: CentraalPlanbureau.

White, H. 1982. Instrumental variables regression with independent observations. *Econometrica* 50: 483–500.

White, H. 1984. *Asymptotic theory for econometricians*. Orlando: Academic.

White, H. 1985. Instrumental variables analogs of generalized least squares estimators. *Journal of Advances in Statistical Computing and Statistical Analysis* 1: 173–227.

Wright, S. 1925. *Corn and Hog correlations*. Washington, DC: US Department of Agriculture, Bulletin 1300.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57: 348–368.

Zellner, A., and H. Theil. 1962. Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica* 30: 54–78.

Instrumentalism and Operationalism

Lawrence A. Boland

Abstract

Instrumentalism and Operationalism are the methodological doctrines associated respectively with Milton Friedman and Paul Samuelson. Each has a long philosophical history. Instrumentalism was the 18th-century doctrine created to deal with the Newton mechanics; Operationalism was the early 20th-century doctrine created to deal with Einstein' general relativity. With Instrumentalism one can say that theories do not have to be true, just

useful – as Friedman argued in 1953. With Operationalism one is required to express theories only in terms of observable and measurable variables. Samuelson's early work was designed to demonstrate how theory can be made operational and thus potentially refutable.

Keywords

Assumptions; Austrian economics; Behaviouralism; Berkeley, Bishop G.; Bridgman, P.; Consumer's demand curve; Einstein, A.; Friedman, M.; Instrumentalism; Mathematics and economics; Methodology of economics; Neoclassical economics; Newton, I.; Operationalism; Operationally meaningful; Perfect competition; Positive economics; Positivism; Psychology; Revealed preference theory; Samuelson, P. A.; Slutsky equation; Tautologies; Utility functions; Verificationism; Weak axiom of revealed preference

JEL Classifications

B4

For any reader familiar with economic literature, the conjunction of the two methodological doctrines of Instrumentalism and Operationalism immediately calls to mind the methodological pronouncements of Milton Friedman (1953) and Paul Samuelson (1947, 1965, 1983). Interestingly, when making their methodological pronouncements, neither Friedman nor Samuelson mentions a philosopher to support his methodological viewpoint or to indicate an inspiration. Nevertheless, each of these methodologies is intended to solve a philosophical problem. Each has a philosophical history.

Friedman's Instrumentalism

Historically, Instrumentalism was a response to the success of Newton's laws of physics that were claimed to explain facts of nature and in particular explain the movements of the planets. Given the success of Isaac Newton's physics, the problem was thought to be that ordinary people would look to science instead of the church for true explanations of nature. Thus, the fear was that any recognized authority of science might undermine faith. Bishop George Berkeley in the early 18th century promoted the idea that we should allow science to be seen *only* as instruments or tools to solve practical problems (for example, instruments to calculate movements of the planets and thus make astronomical predictions). In this way, scientific explanations should not be considered true, only useful. If people would accept the doctrine of Instrumentalism then science and religion could comfortably coexist.

Friedman's Instrumentalism has nothing to do with religion. Instead it is a response to the demands of the critics of neoclassical economics and in particular criticism of the assumptions of perfect competition. In the 1930s and 1940s, the climate of philosophical opinion concerning any claim to scientific knowledge was that it must be realistic in the sense that any scientific claim can be verified as true. With this in mind, critics of neoclassical economics claimed that the assumptions of neoclassical theory would have to be shown to be true if they are claimed to be the basis of true explanations of economic behaviour – or, perhaps more importantly, if they are used to form economic policy such as that involving labour and employment (for example, Lester 1946, 1947). The key question was: must the assumptions of economic theory be true in order to be useful? In his 1953 article 'The Methodology of Positive Economics', Friedman argues that useful theories do not have to be based on true assumptions, or can even, in some cases be based on assumptions that are known to be false.

Few economists today would think any theory should have to be absolutely true for any serious consideration of that theory, whether or not it is to be used for the formation of an economic policy (for example, Aumann 1985). That is, few today think we should be concerned with the question of the absolute truth status of any scientific theory. Instead, all that is hoped is that the theory is the best available as determined by the current scientific conventions. Instrumentalism is an answer to a different question: 'What is the *role* of scientific

theories?' As noted above, Bishop Berkeley had long ago answered this question. Scientific theories should not be considered true or false, simply because, so long as they are useful tools of analysis or prediction, their truth status does not matter. Instrumentalism will never be seen as a satisfactory methodology to economists who think the truth status of economic theory matters (for example, Lawson 1997, 2003).

As it turns out, Instrumentalism can be a very useful tool for the defense against any competing methodological doctrine that claims either that scientific theories are true or that scientific theories must be proven true before they can be used to deal with real practical problems. Proponents of Instrumentalism can always respond to critics by merely claiming that the use of Instrumentalist methodology has proven to be very useful.

Samuelson's Operationalism

Operationalism is usually attributed to physicist Percy Bridgman's 1927 book, although several writers may have taken similar positions before (see Mirowski 1998; Hands 2004). Bridgman's contribution was apparently a response to the growing interest in Albert Einstein's general relativity-based theory of physics that was being seen as the successor to the classical physics of Newton. Einstein's physics was a challenge for most people of the day to understand. Bridgman was seen to be offering a common-sense view of physical theory and methodology that fitted better with what most people understood.

The basic idea of Operationalism is that explanations should be based only on concepts and variables that can be defined by the operations used to measure them. Interestingly, Einstein in 1905 started out trying to explain his theory of special relativity by showing how what we mean by 'simultaneous' is not as obvious as we might think should we try to operationalize it, that is, try to define it in terms of the operations used to measure it. However, he ultimately rejected such an operationalist approach as it made the development of his theory of general relativity paradoxical or impossible (Schilpp 1949).

Apparently, in the 1930s Operationalism was seen as a plausible means of implementing positivism. Supposedly, if every concept used in one's theory can be operationally defined and is thereby observable, then any empirical verification of a scientific theory would be beyond dispute. Moreover, in the 1930s it was commonly believed that scientific theories were meaningful while philosophical or religious theories were not. And the basic notion to support this was that one could verify scientific theories but not philosophical or religious theories. Operationalism was seen by some to be an avenue to make the common verificationist methodological perspective plausible.

About the same time as Bridgman was arguing for Operationalism, there was a movement towards behaviouralism in psychology. The motivation there was that we should avoid constructs such as consciousness or mind and use only observable behaviour. Being observable behaviour means that one could in principle define operations such that one would be able to measure human behaviour. And this is exactly what can be seen in Samuelson's advocacy of Operationalism in economics. What Samuelson wanted to purge from economic theory and in particular from the theory of the consumer was psychology (1938). That is, in Marshallian neoclassical theory, the consumer is thought to be maximizing utility whenever making a decision about what to buy; however, we cannot observe, let alone measure, the level of utility achieved. So how do we know that it is maximized? Do we have to turn our analysis of consumer decisions over to the psychologists? Samuelson thought not. Perhaps his motivation was only the recognition that, if one were to promote mathematical formulations of economics, everything would need to be quantifiable. But the advocacy of Operationalism goes beyond this by requiring the use of only observable variables to derive the fundamental laws of economics, such as the so- called Slutsky equation and the consumer's demand curve. Samuelson claimed that one did not have to assume the existence of a utility function for the consumer but only that the consumer makes well-defined and consistent observable choices. These choices and whether they are consistent is completely and

directly observable and are so without any reference to psychology or utility.

Samuelson went on in his Ph.D. thesis (published in 1947; see also 1998) to say that it is possible to construct all of the important ideas in economics in such a manner that they can be shown to be in principle falsifiable. He called empirically falsifiable statements 'operationally meaningful statements'. It should be noted, however, that by 1947 he was no longer taking the extreme view taken in 1938, which required that all assumptions of a theory be directly observable, but instead said that a theory need only be shown to have implications that are falsifiable with observable data. Any theory or model that has such implications would henceforth be deemed to be 'operationally meaningful'.

This invocation of Operationalism is somewhat suspect. At the time Samuelson was promoting the use of mathematics in economics, critics, particularly some of the Austrian School, were claiming that all mathematical propositions are tautologies (see Hutchison 1935, 1938). Samuelson, by requiring any theory or model to be 'operationally meaningful' (that is, falsifiable), was really just avoiding the critics, for there is no conceivable evidence that can refute a tautology. So, when he showed that falsifiable statements can be derived from his mathematical model of an economic theory, he proved that his mathematical model is not a tautology. Thus, Operationalism offered a means of dealing with the critics of the use of mathematics in economics.

The Methodological Failures of Friedman and Samuelson

By 1950 Samuelson had to admit that his original operationalist programme to purge psychological concepts such as utility from consumer theory was a failure. Consistency of today's choice with past choices presumes no change in tastes. When he revisited his version of consumer theory in 1948, he introduced preferences into the discussion. His key assumption was to be called the weak axiom of revealed preference. However, once it was recognized that the weak axiom was not sufficient for the purposes of consumer theory, the introduction of a strong version revealed that his revealed preference analysis was logically equivalent to old fashioned utility- based analysis (see Wong 1978, 2006). This undermined any further interest in promoting Operationalism in economics.

Friedman's version of Instrumentalism begs many questions. Who decides what is meant by 'useful' or which empirical facts need to be predicted with one's economic model? What one economist might think an obviously successful policy that justified the use of obviously false assumptions might not be accepted as successful by critics of the realism of those assumptions. So some may think that Instrumentalism can be used to defend its use in a self-referential form by economists, others can just as easily say that the primary and perhaps sole use of Instrumentalism is to avoid criticism of theories and hence of policies recommended on the basis of those theories. Instrumentalism still lives in economics, although in a somewhat muted form (see Boland 2003, Chaps. 4 and 5). The dominance of formal mathematical modelling has often been supported by appeals to Instrumentalism (for example, Aumann 1985, pp. 31–2). Rarely, however, will a promoter of mathematics in economics refer to Friedman's essay, but the methodological position taken is the same.

See Also

- ▶ Assumptions Controversy
- ▶ Conventionalism
- ▶ Friedman, Milton (1912–2006)
- ▶ Samuelson, Paul Anthony (1915–2009)

Bibliography

Aumann, R. 1985. What is game theory trying to accomplish? In *Frontiers of economics*, ed. K. Arrow and S. Honkapohja. Oxford: Basil Blackwell.

Boland, L. 1979. A critique of Friedman's critics. *Journal of Economic Literature* 17: 503–522.

Boland, L. 2003. *The foundations of economic method: A Popperian perspective*. London: Routledge.

Bridgman, P. 1927. *The logic of modern physics*. New York: Macmillan.

Friedman, M. 1953. Methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.
Hands, D.W. 2004. On operationalisms and economics. *Journal of Economic Issues* 38: 953–968.
Hutchison, T. 1935. A note on tautologies and the nature of economic theory. *Review of Economic Studies* 2: 159–161.
Hutchison, T. 1938. *The significance and basic postulates of economic theory*. London: Macmillan.
Lawson, T. 1997. *Economics and reality*. London: Routledge.
Lawson, T. 2003. *Reorienting economics*. London: Routledge.
Lester, R. 1946. Shortcomings of marginal analysis for wage-employment problems. *American Economic Review* 36: 63–82.
Lester, R. 1947. Marginalism, minimum wages, and labor markets. *American Economic Review* 37: 135–148.
Mirowski, P. 1998. Operationalism. In *Handbook of economic methodology*, ed. D.W. Hands, J. Davis, and U. Mäki. Cheltenham: Edward Elgar.
Samuelson, P. 1938. A note on the pure theory of consumer behaviour. *Economica* 5(n.s.): 61–71.
Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge: Harvard University Press.
Samuelson, P. 1948. Consumption theory in terms of revealed preference. *Economica* 15(n.s.): 243–253.
Samuelson, P. 1950. The problem of integrability in utility theory. *Economica* 17(n.s.): 355–385.
Samuelson, P. 1965. *Foundations of economic analysis*, 2nd ed. New York: Atheneum.
Samuelson, P. 1983. *Foundations of economic analysis*, 3rd ed. Cambridge, MA: Harvard University Press.
Samuelson, P. 1998. How *foundations* came to be. *Journal of Economic Literature* 36: 137–186.
Schilpp, P.A. (ed.). 1949. *Albert Einstein: Philosopher-scientist*. Evanston: Library of Living Philosophers.
Wong, S. 1973. The 'F-twist' and the methodology of Paul Samuelson. *American Economic Review* 63: 312–325.
Wong, S. 1978. *The foundations of Paul Samuelson's revealed preference theory*. London: Routledge & Kegan Paul.
Wong, S. 2006. *The foundations of Paul Samuelson's revealed preference theory*, 2nd ed. London: Routledge.

Insurance

J. J. McCall

Insurance is an ancient institution. It is impossible to reflect on evolutionary processes without recognizing the intrinsic role of insurance. Any species that relied on nature's harmony and regularity and ignored its stochastic whims was soon extinct. The position adopted here is that uncertainty is one of the decisive determinants of individual behaviour. 'Individual' includes not only early and modern man, but also plants and animals. Furthermore, the response to uncertainty is both adaptive and dynamic. Martingale models are ideally suited to portray these responses. The goal of the individual is to maximize expected utility and for early man this was roughly equivalent to maximizing the probability of survival. In order to achieve this goal, he devised a variety of insurance mechanisms he established institutions that included a large element of flexibility so they could readily adjust to nature's stochastic quirks. Insurance and the ensuing flexibility were key components of his decisionmaking. Aggregating these individual responses over the entire group reveals one critical aspect of the society's culture. The individual quickly perceived that some of the most effective devices for mitigating uncertainty entailed *cooperative arrangements*.

At the same time that individuals and bands are designing mechanisms for coping with a fluctuating environment, nature is inexorably monitoring these activities and eliminating those individuals and bands who respond too slowly to adversity or are unlucky, and despite their best efforts, are overcome by misfortune. These Job-like extinctions are more likely in harsh and highly variable environments. We expect that those who survive in these environments are quite distinct from survivors where nature is relatively benevolent. The excellent study by Minnis (1985) indicates how persistent misfortune affects existing social institutions and induces significant modifications. Minnis examined three periods of environmental stress and shows the enhanced survival value of innovation and flexibility.

Insurance also manifests itself in unconscious natural selection. Lowell (1985) has shown that as the predictability of the environment decreases, the safety factory of the biological structure increases. Roughly speaking, the organism adapts to the distribution of the maximum stress, rather than to the distribution of the actual stress. In other

words, the difference between the stress-capability of the organism and the actual stress encountered, measures the 'slack'. The presence of 'slack' or flexibility enhances the survivability of the organism. Thus a fairly conservative way of handling unanticipated shocks is to imitate nature and use the distributions of the extremes (maximum period of drought or minimum density of prey). Leadbetter et al. (1983) is a fine, comprehensive treatment of extreme value distributions. Flocking (Morse 1970) is another of the almost endless variety of insurance mechanisms that evolved in response to the threat of extinction. The study of extinction is a discipline in itself. A good sampling of recent research is the book edited by Nitecki (1984) with Diamond's paper on isolated populations an outstanding contribution. Martin's piece on catastrophic extinctions is the most pertinent. Also see Mangel and Ludwig (1977) for a stochastic analysis of extinction in competitive struggles among species. The classic studies by Slobodkin (1961) and MacArthur (1972) are illuminating. An elegant model of extinction could be devised using the contact process methods in Liggett (1985).

The Basic Elements of Insurance

We have already described the essential aspects of insurance. The remaining task is to translate them into the modern language of economics and probability. The original research on the economics of insurance was conducted by Arrow (1971) and Borch (1968). Hirshleifer and Riley (1979) contains a fine survey of insurance; also see Lippman and McCall (1982). The neatest presentation of stochastic dominance in an insurance setting is the paper by Lippman (1972) that has gone unnoticed in spite of its excellence.

Individuals in modern societies are unable to predict the time and magnitude of events that profoundly affect their well-being. Insurance, in all its guises, is the institution that mitigates the influence of uncertainty. The individual invests in a host of activities *now* to insure that the timing and magnitude of unfortunate future events will be less harmful. These activities enable firms and individuals to trade risks among themselves. The most familiar of these transfers is the ordinary insurance contract. The essence of this contract is the payment of a fee by the insuree in exchange for the insurer's promise to pay a certain sum of money provided a stipulated event occurs.

One of the simplest insurance contracts has the following structure. The insured pays a fixed premium y to avoid the small probability p of incurring L, a large loss. For simplicity, ignore loading charges (i.e. the charge to cover the administrative costs associated with writing and overseeing the insurance contract), and set this premium equal to the actuarial value of the loss plus an amount c, the compensation to the insurer for assuming the risk:

$$y = pL + (1-p)0 + c = pL + c. \quad (1)$$

Such a policy is advantageous to both the insured and insurer. The insured possesses a concave utility function and is therefore eager to pay y to dispense with this risk. The insurance company is able to pool independent risks and via the law of large numbers converts risky contracts into almost 'sure' things.

Because of its fundamental importance we state a simple version of the law of large numbers, namely, the weak law of large numbers for a fair Bernoulli random variable. (This version applies when $L = 1$ and $p = 1/2$.) A fair coin is flipped n times. Let the random variable X_n be given by

$$X_n = \begin{cases} 1, & \text{if a head occurs on } n\text{th flip} \\ 0, & \text{if a tail occurs on the } n\text{th flip,} \end{cases}$$

and S_n is defined by

$$S_n = \sum_{i=1}^{n} X_i.$$

The proportion of heads in the first n trials is simply S_n/n.

At first one might think that the probability of exactly n heads in $2n$ trials should be high because $p = 1/2$, but

$$P(S_{2n} = n) = \binom{2n}{n} 2^{-2n} \simeq 1/\sqrt{\pi n}. \quad (2)$$

Thus, the probability of exactly n heads in $2n$ trials goes to zero as n gets large.

The weak law of large numbers states that the probability of $S_n|n$ deviating from $\frac{1}{2}$ by any fixed positive quantity goes to zero as n goes to infinity. More precisely,

$$\lim_{n\to 0} P\left(\left|S_n/n - \frac{1}{2}\right| > \varepsilon\right) = 0, \quad \text{for any } \varepsilon > 0. \quad (3)$$

Proof Letting $\widehat{\Sigma}$ designate the sum over the set of integers k such that $|k/n - \frac{1}{2}| > \varepsilon$, we have

$$\begin{aligned} P\left(\left|S_n/n - \frac{1}{2}\right| > \varepsilon\right) &< \widehat{\Sigma}\left[\left(k/n - \frac{1}{2}\right)/\varepsilon\right]^2 \binom{n}{k} \\ \times 2^{-n} &\le \varepsilon^{-2} \sum_{k=0}^{n} \left(k/n - \frac{1}{2}\right) 2 \binom{n}{k} 2^{-n} \\ &= \varepsilon^{-2} \mathrm{Var}(S_n/n) = p(1-p)/n\varepsilon^2 \\ &\le 1/4n\varepsilon^2. \end{aligned}$$

where the first inequality follows from the fact that $\left(k/n - \frac{1}{2}\right)^2 > \varepsilon^2$, the second inequality from the fact that additional non-negative terms are being summed, and the final inequality from the fact that $x(1-x)$ is maximized at $x = \frac{1}{2}$. Q.E.D.

Note that the law of large numbers does not say that if you and I play a game of chance with a fair coin, I will lead approximately half the time. In fact, if I win the game, I am likely to have led for most of the game.

By this law S_n/n converges in probability to $\frac{1}{2}$. Thus, an insurance company has considerable control over its average loss.

When contracting (risk transfer) transpires in an uncertain environment, two basic problems present themselves: *moral hazard* and *adverse selection*. Both are founded on imperfect information. The prototypical example of these problems is the insurance contract between an insurance company (the principal) and the insured (the agent). By paying a premium the agent transfers the risk associated with a particular activity to the principal. This risk transfer affects the incentives and behaviour of the agent. It is these incentive effects that are commonly referred to as the moral hazard problem. It has its roots in the inability of the principal to costlessly observe the actions of the agent. Hence when the untoward event occurs, the principal is not sure whether it was caused by the agent's carelessness or by chance. Moral hazard can be reduced by requiring the agent to bear some of the costs of the contingency and/or by monitoring the agent's behaviour. Adverse selection is similar to moral hazard in that the problem arises because the principal does not have costless access to information possessed by agents and vice versa. For example, some purchasers of health insurance have more information about their health status than insurance companies. Because the insurance company cannot discriminate perfectly between healthy and sickly agents, the latter will pose as healthy agents and be 'adversely selected' (insured) by the principal. Insurance companies can and do cope with these informational asymmetries by (a) experience rating, that is, continually adjusting premiums to reflect the size and incidence of each agent's claims and (b) designing policies that elicit the information necessary for partitioning agents into distinct categories.

These problems have received an enormous amount of study by economic theorists. Much of the analysis is static and, of course, gives rise to many perplexes. When properly formulated as dynamic stochastic control problems the perplexities diminish and the solutions accord with those that have been used for centuries by business firms.

Alternative Insurance Mechanisms

The insurance contract is only one of a multitude of devices that have been created for coping with the risks that afflict any economic system. These risks include not only fire, theft, sickness, and death but also fluctuating prices, equipment malfunctions, zero inventory levels causing unsatisfied demands, and failure of basic research ranging from falsely 'proved' theorems to unisolated viruses. The existence of futures

contracts permits the farmer or food processor to specialize in production, while the speculator specializes in risk-bearing. The risks of equipment failure can be reduced by improved design and maintenance procedures like redundancy and frequent inspection. The probability of an unfulfilled demand can be diminished by maintenance of larger inventories. The costs of research failure are frequently insured against by initiation of a large number of relatively independent projects (self-insurance), or, where the costs are large and uncertain, by adoption of inefficient contractual procedures like the cost-plus, fixed-fee contract (government insurance).

The basic institution for shifting the risks of business from entrepreneurs to the general public is the securities market. Individuals can diversify their portfolio of stocks to achieve an acceptable level of expected return for a given level of risk. This ability of individuals to spread risks thereby permits firms to engage in projects which otherwise would be unacceptable. Consequently, society is better off.

These insurance arrangements are, however, far from ideal. It is usually impossible for a firm to transfer *only* rights to the outcomes of its highly risky ventures. In contrast with the futures market, the stock market is usually incapable of separating production and risk, leaving the former to the entrepreneur and transferring the latter to the general public. Instead, the stock certificate is a relatively blunt instrument for disentangling risk and production. The fact that society has not created a sharper instrument attests to the refractory nature of this problem.

The Probability of Bankruptcy

The goal of many firms and indeed of many organisms is to maximize the probability of survival.

Consider the following simple version of the bankruptcy problem. An entrepreneur begins with wealth $W_0 = 0$ and acquires \$1 with probability p and loses \$1 with probability $1-p$. If he reaches $\$ - b$ before he reaches a wealth of \$a, he is bankrupt; whereas if he reaches a before $-b$, he can issue stock and essentially reduce his ruin probability to zero. Let z be the probability of hitting a before $-b$, and Y_n the position of the entrepreneur after n periods (trials). Then

$$X_n = \left(\frac{1-p}{p}\right)^{Y_n}$$

is a martingale with $E(X_n) = 1$, all n. Hence, the probability of survival, z, is the solution to:

$$z\left(\frac{1-p}{p}\right)^{a} + (1-z)\left(\frac{1-p}{p}\right)^{-b} = 1$$

which implies that

$$z = \frac{b}{a+b}.$$

When survival is the criterion function in a finite horizon problem, it is easy to show that the optimal policy is conservative (timid) for the first few trials (the entrepreneur has control over the size of the bet at each trial), but becomes bold as the horizon approaches and success has not yet been achieved. Using standard terminology, the stochastic behaviour of the entrepreneur would switch from risk aversion to risk preference at some critical point. This problem is solved in Lippman and McCall (1980) and Houston and McNamara (1985).

See Also

- Adverse Selection
- Life Insurance
- Moral Hazard
- Risk
- Uncertainty

Bibliography

Arrow, K.J. 1964. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31, April: 91–96.

Arrow, K.J. 1971. *Essays in the theory of risk-bearing*. Chicago: Markham.

Borch, K.H. 1968. *The economics of uncertainty*. Princeton: Princeton University Press.
Bühlmann, H. 1970. *Mathematical methods in risk theory*. Berlin/Heidelberg/New York: Springer.
Dubins, L., and L.J. Savage. 1965. *How to gamble if you must*. New York: McGraw-Hill.
Ehrlich, I., and G.S. Becker. 1972, July–August. Market insurance, self-insurance, and self-protection. *Journal of Political Economy* 80(4): 623–648.
Hirshleifer, J. and J.G. Riley. 1979, December. The analytics of uncertainty and information – An expository survey. *Journal of Economic Literature* 17(4): 1375–1421.
Houston, A., and J. McNamara. 1985. The choice of prey types that minimizes the probability of starvation. *Behavioral Ecology and Sociology* 17: 135–141.
Leadbetter, M.R., G. Lindgren, and H. Rootzen. 1983. *Extremes and related properties of random sequences and processes*. New York: Springer.
Liggett, T.M. 1985. *Interacting particle systems*. New York: Springer.
Lippman, S.A. 1972. Optimal insurance. *Journal of Financial Quantitative Analysis* 7: 2151–2155.
Lippman, S.A., and J.J. McCall. 1980. Constant absolute risk aversion, bankruptcy, and wealth-dependent decisions. *Journal of Business* 53: 285–296.
Lippman, S.A., and J.J. McCall. 1982. The economics of uncertainty: Selected topics and probabilistic methods. In *Handbook of mathematical economics*, ed. K.J. Arrow and M. Intriligator. Amsterdam/New York: North-Holland.
Lowell, R.B. 1985. Selection for increased safety factors of biological structures as environmental unpredictability increases. *Science* 228: 1009–1011.
MacArthur, R.H. 1972. *Geographical ecology*. Princeton: Princeton University Press. Reprinted, 1984.
Mangel, M., and D. Ludwig. 1977. Probability of extinction in a stochastic competition. *SIAM Journal of Applied Mathematics* 33: 256–267.
Minnis, P.E. 1985. *Social adaptation to food stress*. Chicago: University of Chicago Press.
Morse, D.H. 1970. Ecological aspects of some mixed species foraging flocks of birds. *Ecological Monographs* 40: 119–168.
Nitecki, M.H. 1984. *Extinctions*. Chicago: University of Chicago Press.
Pauly, M.V. 1968. The economics of moral hazard: comment. *American Economic Review* 58, June: 531–536.
Pratt, J.W. 1964. Risk aversion in the small and in the large. *Econometrica* 32: 122–136.
Radner, R. 1981. Monitoring cooperative agreements in a repeated principal-agent relationship. *Econometrica* 49: 1127–1148.
Riley, J.G. 1975, April. Competitive signalling. *Journal of Economic Theory* 10(2): 174–186.
Riley, J.G. 1979, March. Information equilibrium. *Econometrica* 47(2): 331–359.
Rothschild, M., and J.E. Stiglitz. 1976, November. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 90(4): 629–649.
Rubinstein, A., and M.E. Yaari. 1983. Repeated insurance contracts and moral hazard. *Journal of Economic Theory* 30: 74–97.
Seal, H.L. 1969. *Stochastic theory of a risk business*. New York: Wiley.
Slobodkin, L.B. 1961. *Growth and regulation of animal populations*. New York: Holt, Rinehart & Winston.
Spence, A.M. 1974. *Market signalling: Informational transfer in hiring and related processes*. Cambridge, MA: Harvard University Press.
Zeckhauser, R. 1970, March. Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* 2(1): 10–26.

Insurance Mathematics

James C. Hickman and Edward W. Frees

Abstract

Insurance mathematics is concerned with the valuation of obligations arising from insurance contracts. At contract initiation, valuation is known as premium determination or ratemaking, whereas, for a contract already in force, valuation is known as reserve determination. Updating these values as information is revealed involves important techniques known as experience adjustment. Models of insurance mathematics are based on probability theory and financial economics. These models are calibrated with insurance experience and present values from returns on investments in asset markets.

Keywords

Benefit premium; Calibration; Central limit theorems; Collective risk theory; Compound interest; Continuous interest rate; Defined benefits; Equivalence principle; Expected values; Health insurance; Insurance mathematics; Liability; Life insurance; Mortality; Pensions; Portfolio theory; Present value; Probability density function; Recursion relationships;

Risk management; Risk theory; Selection bias; Workers' compensation insurance

JEL Classifications
G22

Mathematics and insurance have developed along parallel paths during the past 350 years. It is difficult to identify an economic activity more closely tied to mathematics than insurance. Since the genesis of probability ideas in the mid-17th century, there have been times when mathematical developments were ahead of insurance practice. At other times, commercial necessity required improvisations that did not rest on solid mathematical foundations. In general the science and the application moved together.

Reserves and Premiums: Long-Term Coverages

Two related valuation problems are to establish a price, or premium, and to estimate the liability created by a contract. The basic tools for solving these problems are expected values and compound interest combined with an economic concept, the equivalence principle. The equivalence principle requires, at the time the coverage is activated, that the expected present value of premiums equals to expected present value of benefits. Following the issuance of the coverage, the principle can be extended to define the liability of the insurer as the expected present value of future benefits less the expected present value of future premiums.

Long-term insurance contracts have the possibility of extending for many years. The time of benefit payment, and for some contracts the amount of payments, often depend on the length of the survival time of the insured. Specifically, let T denote the random variable time until death. One example of a long-term coverage is life insurance with a single payment of benefits at time T Another example is a life annuity with many payments of benefits paid during survival, up to time T The life insurance model would apply to financing the replacement of equipment from light bulbs to generators. The mathematics of annuities would apply to funding equipment maintenance costs.

To illustrate, consider a life insurance policy paying a benefit b at death to be funded by a premium π, paid at a continuous annual rate until death. For the time until death random variable, let $\mathrm{s}(t) = \Pr(T > t)$ be the survival function. Then the equivalence principle determines the premium π by the equation

$$-b\int_0^\infty e^{-\delta t}\mathrm{s}'(t)dt = \pi \int_0^\infty e^{-\delta t}\mathrm{s}(t)dt. \qquad (1)$$

Here, $-\mathrm{s}'(t)$ is the probability density function of time until death and δ is the continuous interest rate, also called the force of interest. It will be assumed constant for simplicity. It is defined through the relation $1 + i = e^{\delta}$, where i is the annual effective rate of interest. The premium rate π is known as a 'benefit premium'; it is computed assuming that firms are risk neutral and that there are no transactions costs. In commercial practice, the benefit premium π will be increased to a contract premium G , $G > \pi$. The contract premium will contain provisions for expenses, profits and risk. The equivalence principle can be extended to include these elements.

The liability of the insurer, denoted by ${}_sV$ given survival to s , $s \geq 0$, would be given by the equivalence principle as

$$\begin{aligned} {}_sV + \pi\int_s^\infty e^{-\delta(t-s)}\mathrm{s}(t-s|\, s<T)dt \\ = -b\int_s^\infty e^{-\delta(t-s)}\mathrm{s}'(t-s|\, s<T)dt. \end{aligned} \qquad (2)$$

In words, the liability is the expected present value, also called the actuarial present value, of future benefits less the actuarial present value of future premiums. In this equation, the conditional survivorship function is $\mathrm{s}(t-s|\, s<T) = \mathrm{s}(t)/\mathrm{s}(s)$.

In Eq. (1) for the premium rate and Eq. (2) for the reserves, making benefits and premiums a function of time survived, $\mathrm{b}(t)$ and $\pi(t)$, creates no conflict with the equivalence principle. There are practical reasons for requiring ${}_sV \geq 0$. This prevents a voluntarily withdrawing insured from

leaving the insurer with a negative liability, a non-collectable asset.

As another special case, we now consider a life annuity with benefits at an annual rate b starting at retirement time r, funded by a continuously paid premium rate π paid during survival to r. Such a contact would be a building block of a pension plan. The equivalence principle yields

$$\pi \int_0^r e^{-\delta t} \mathrm{s}(t) dt = b \int_r^\infty e^{-\delta t} \mathrm{s}(t) dt, \qquad (3)$$

allowing us to compute the premium rate π based on survivorship and interest information.

To illustrate how other contracts can be accommodated, we consider the life annuity case that also includes a so-called 'return of premiums' feature. With this feature, there is an additional benefit consisting of the accumulated premiums (with interest) that are paid at death before time r. The benefit side of formula (3) is increased by

$$\begin{aligned} &-\pi \int_0^r \left(\int_0^t e^{\delta x} dx \right) e^{-\delta t} \mathrm{s}'(t) dt \\ &= -\pi \left(\mathrm{s}(r) \int_0^r e^{-\delta t} dt + \int_0^r e^{-\delta t} \mathrm{s}(t) dt \right), \end{aligned} \qquad (4)$$

where the right-hand side is from an integration by parts. With this additional benefit, from formula (3) we have

$$\pi \mathrm{s}(r) \int_0^r e^{-\delta t} dt = b \int_r^\infty e^{-\delta t} \mathrm{s}(t) dt,$$

a result that might have been derived by general reasoning from the equivalence principle.

We return to the life annuity premium displayed in formula (3). The equivalence principle yields a reserve liability at time s, $0 \le s$, of

$$\begin{aligned} &sV + \pi \textstyle\int_s^r e^{-\delta(t-s)} \mathrm{s}(t-s \,|\, s < T) dt \\ &= b \textstyle\int_s^\infty e^{-\delta(t-s)} \mathrm{s}(t-s \,|\, s < T) dt \; 0 < s < r_s V \\ &= b \textstyle\int_s^\infty e^{-\delta(t-s)} \mathrm{s}(t-s \,|\, s < T) dt \; r < s. \end{aligned}$$

The key role played by the survival function and the assumed interest rate in these typical formulas is clear.

Reserves and Premiums: Short-Term Coverages

Short-term coverages include most individual property/casualty, health and group insurance policies. They are characterized by the reduced role of present values. In addition, the benefit amount is typically a random variable. Its value will depend in health insurance on the services provided, and in property insurance on the extent of the property damage. Premiums and reserves will continue to be determined by the equivalence principle. In the time period between the occurrence of a loss event and its settlement, available information about the loss event will determine reserve amounts.

The expected value of benefit payments for short-term coverages is given by

$$\begin{aligned} \pi &= \mathrm{E}\left(\sum_{i=1}^N X_i \right) = \mathrm{E}\left(\mathrm{E}\left(\sum_{i=1}^N X_i \right) | N = n \right) \\ &= \mu \mathrm{E}(N), \end{aligned}$$

where N denotes the random number of losses during the insurance period, X_i is the loss amount arising from loss i and $\mathrm{E}\, X_i = \mu$.

If the distribution of N is Poisson and N and the loss amounts are independent, then

$$S = X_1 + \ldots + X_N \qquad (5)$$

has a compound Poisson distribution. Clearly many distributions, such as the binomial or negative binomial, could be used for the distribution of N.

The reserve liability for short-term coverages uses information about loss events and the loss reserve is

$$\mathrm{E}(X_1 + \ldots + X_N | N = n) = n\mathrm{E}(X) = n\mu$$

and n is the number of losses incurred.

Risk theory is the study of the distribution of total losses and the management of their inconvenient consequences. The earliest contributions to risk theory build on the model for long-term coverages. We start with loss variables

$$L_j = b_j e^{-\delta T_i} - \pi_j \int_0^{T_i} e^{-\delta s} ds,$$

and study the distribution of $S = L_1 + \ldots + L_n$. Here, T_i is a random variable representing the future lifetime of an individual. This study is known as individual risk theory because the variable S is based on n individual loss variables. If the loss variables are assumed to be mutually independent, then

$$Z = \frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}}$$

will have, as a result of an extension of the central limit theorem, an approximate normal distribution, with mean zero and variance one. In contrast, the direct study of the distribution of S as in formula (5) is called collective risk theory. Approximating the distribution of S has been an active topic in actuarial research since early in the 20th century.

Experience Adjustment: Long-Term Coverages

Valuation of long-term coverages requires assumptions about the realizations about interest rates and mortality in the distant future. In this dynamic world it is almost certain that the results expected by an insurance system will not be obtained. For many contracts, it has become customary for insurers to make assumptions that many financial analysts would view as conservative for pricing at contract initiation. As better than anticipated experience is realized, excess funds are realized that can be directed to the insured in a mutual insurance organization or to owners of the insurance company. For the insured, these are additional (non-contractual) benefits; depending on the regulatory environment, these additional benefits come in the form of dividends or bonuses.

Reconciling anticipated to actual experience is done periodically, not just at the conclusion of the contract. Because of this periodic reconciliation, recursion relationships are important tools for measuring and adjusting for deviations from expected results.

Specifically, let ${}_{s-1}F$ be the fund, possibly the insurance reserve, at the end of policy year s-l. Define ${}_sP$ to be the premium paid at the beginning of policy year s , $\mathrm{E}({}_sB)$ the expected benefits paid at the end of policy year s and ${}_sF$ the expected fund at the end of policy year s. We simplify and assume that $E({}_sB) = b\, q_s$. A basic recursive relationship is

$$({}_{s-1}F + {}_sP)(1+i) - bq_s = {}_sF\, p_s, \qquad (5a)$$

where i is the expected annual interest rate and $p_s = 1 - q_s$. Formula (5) can be written as

$$({}_{s-1}F + {}_sP)(1+i) - q_s(b - {}_sF) = {}_sF. \qquad (6)$$

If the actual experience yields i' and q_s', then formula (6) can be written as

$$({}_{s-1}F + {}_sP)(1+i') - q_s'(b - {}_sF) = {}_sF + D, \qquad (7)$$

where D is a deviation of actual from expected results. If $D > 0$, the amount might be paid to the insured in a mutual insurance organization or to owners of the insurance company.

Subtracting formula (6) from (7), yields

$$D = (q_s - q_s')(b - {}_sF) + (i' - i)({}_{s-1}F + {}_sP). \qquad (8)$$

The first term on the right-hand side of formula (8) is called the mortality contribution and the second term the interest contribution. Formulae used in practice also contain a term for the difference between expense loading and actual expenses.

To study life annuities, the general formula (6) can be modified during the benefit payment period to yield

$${}_{s-1}F(1+i) - p_S b = p_{SS} F. \qquad (9)$$

Replacing the expected parameters with experience parameters, we have

$$ {}_{s-1}F(1+i') - p'_s b = p'_s({}_sF + D). \qquad (10) $$

Subtracting formula (9) from formula (10) yields

$$ {}_{s-1}F(i'-i) + \left(p_s - p'_s\right)(b + {}_sF) = p'_s D. $$

If $D > 0$, this expression could be the basis of a dividend to surviving annuitants.

These recursion relationships are also the basis for flexible coverages where premiums and benefits can be changed by the insured within contractual limits.

Experience Adjustment: Short-Term Coverages

In the first decade of the 20th century industrial accidents were a leading cause of death, a source of much litigation and a major social concern. The advent of workers' compensation insurance replaced litigation with a system based on defined benefits. Employers, in most cases, were required by statute to provide workers' compensation benefits. Because of great variation in the hazards faced in different industries and the lack of loss statistics, initial premiums were set by judgement. The goal was to develop a self-correcting rate estimation process that would also provide incentives to employers to improve industrial safety.

The solution came from the formula

$$ \begin{aligned} (\textit{New rate}) = {} & Z(n) \\ & \times (\textit{observed average losses}) \\ & + [1 - Z(n)] \times (\textit{Initial Rate}), \end{aligned} $$

where the credibility factor is $Z(n)$, n a measure of exposure and $0 \le Z(n) \le 1$. To provide intuition, consider the case where $Z(n) = 1$, known as the 'full credibility' case. Here, the employer's next period premium would consist entirely of observed average losses from the prior period. If the employer had introduced practices to improve industrial safety then this would be reflected in a lower premium. In contrast, consider the case where $Z(n) = 0$. Here, the premium would consist of an initial *rate* that presumably would reflect industry results but not the employer's actual experience. The case $Z(n) = 0$ is the standard for individual coverages. Many employers would fall in the intermediate case, $0 < Z(n) < 1$, known as 'partial credibility'. Premiums for employers in this category would reflect their own industrial safety records as well as benefit from the pooling of risks within an industry.

For the credibility factor, one typically requires $Z'(n) > 0$ and $Z''(n) < 0$. Thus, other things equal, employers with larger exposure (n) enjoy larger credibility but the rate of increase decreases with exposure. A typical credibility function is of the form $Z(n) = n/(n+k)$, $k > 0$. The establishment of k with a satisfactory intellectual foundation has come from Bayesian statistics after its introduction into practice. The credibility idea for experience adjustments is now used in many short-term coverages.

Another type of insurance plan available for groups is known as 'stop-loss' or 'excess of loss' coverage. Large group insurance plans, usually based on employee groups, have distinctly different risk characteristics from individual policies. The sponsor, usually a large organization, is typically willing and able to absorb some variation in benefit payments. Only large and unexpected payments are financially inconvenient to the sponsor. The insurance company is paid to adjudicate and pay benefit claims, and to absorb large and inconvenient benefit payments. Typically, the sponsor maintains an internal account of losses known as an 'experience account'. This account records premiums as income and losses and expenses as expenditures.

We let X be the losses in an experience period and d be the stop loss amount (or d for 'deductible'). The experience account is charged for losses up to d. If $X > d$, then X-d is not charged to the experience account of the sponsor. A risk premium for this experience adjustment is charged on the basis of

$$ \int_d^\infty (x-d) f(x) dx. $$

Model Calibration: Experience Studies

The models introduced suggest that extensive work must be done in estimating survival functions in implementing long-term insurance models. For short-term models, the distribution of the number of losses N, per policy period, and the distribution of X, the loss amount, must be estimated. These efforts are in most applications special cases of statistical estimation.

These estimation projects are generally observational studies. The data come from insurance experience and the subjects have purchased insurance or gained insurance as an employee benefit. The use of general population statistics for insurance purposes has hazards because of potential biases. To illustrate, when studying annuitant mortality, it is well-known that mortality is substantially lower than the general population mortality. This is a selection bias issue; seldom do those in substandard health purchase a life annuity.

Rapid increases in the cost of health services and jury awards in some areas have increased the need to estimate time trends for the distribution of X, loss costs. Because of longer settlement time in some coverages, this estimation has become a major project in loss reserve determination. The rate of increase in health care costs in recent years has been such that estimates of the distribution of X, benefit amount random variable, using information from previous years would result in a distribution significantly to the left of the distribution for the current year. The rate of growth of health care costs is the most important single pricing decision for health insurance.

Model Calibration: Classification

The distributions that enter insurance models are all conditional distributions. Clearly, the distribution of X, loss amount, depends on the time, location and other facts surrounding the insurance loss incident. The distribution of T, time until death, in life insurance depends on a set of classification variables. The purpose of observing these classification variables is to increase the likelihood that the assumed distribution of T will be approximately realized.

The selection of these classification variables may be constrained by law and expense. For example, a determination of the degree of aggression of an applicant for automobile insurance might have a significant impact on the distribution of N, but the expense of collecting the information might be greater than its value in reducing variability.

Model calibration: financial economics

The critical role played by the force of interest δ in premiums and reserves for long-term coverage is clear. The use of an assumed force of interest for an extended period of time will lead, according to common experience, to serious deviations between actual and expected results. Options for moderating these deviations are numerous.

- A statistical model for the force of interest, estimated from past data, could be constructed and the equivalence principle extended to take expectations over the joint distribution of δ_T and T. The joint distribution might also be used to fix an interest rate risk loading into premiums to minimize the inconvenient consequence of variations in δ.
- Arrange the timing of investment cash flows to approximately match the expected cash flows from the insurance operations.
- Pass variations in interest earnings directly to the policy owner as indicated in the section on Experience adjustment, long-term coverages. The insured's account F would absorb variation in investment earnings.
- Use a program of financial derivative contracts to stabilize, for a price, variations in investment income.

Financial economics has not only enriched insurance mathematics by providing risk management tools for the investment risk in conventional insurance contracts, it has also created the possibility of absorbing many traditional insurance risks into special securities traded in worldwide

investment markets. The idea is to use the capital in investment markets, and not just the capital held by insurance companies, to manage risk.

The idea of special securities with contractual payments that approximately match payments from an insurance system has already been developed for several coverages –for example, catastrophe bonds with modified payments following a catastrophe, fitting the definition of the security. A second example is a survivorship bond with regular coupon payments proportional to the number of survivors in a defined group. Such bonds could spread the risk if mortality improvement exceeds the capacity of the sponsor of a pension system.

The market for such special securities is determined, in part, from ideas in financial economics. Portfolio theory would predict that investors would seek securities that have cash flows that are not positively correlated with the regular business cycle. Tying security payments to natural disasters, such as earthquakes and hurricanes, might achieve the sought for independence.

See Also

- ▶ Health Insurance, Economics of
- ▶ Liability for Accidents
- ▶ Life Tables
- ▶ Mortality
- ▶ Pensions
- ▶ Present Value
- ▶ Social Insurance

Bibliography

Booth, P., R. Chadburn, D. Cooper, S. Haberman, and D. James. 1999. *Modern actuarial theory and practice*. London: Chapman and Hall/CRC.

Bowers, N., H. Gerber, J. Hickman, D. Jones, and C. Nesbitt. 1997. *Actuarial mathematics*. Schaumburg: Society of Actuaries.

Klugman, S., H. Panjer, and G. Willmot. 2004. *Loss models: From data to decisions*. 2nd ed. New York: Wiley.

Panjer, H., ed. 1998. *Financial economics: With applications to investments, insurance and pensions*. Schaumburg: Actuarial Foundation.

Intangible Capital

Daniel E. Sichel

Abstract

Intangible capital has played an increasingly important role in economic growth, although firm-level financial and national income accounting practices provide little information about intangibles and do not count many purchases of intangible capital as investment.

Keywords

Economic growth; Financial accounting; Financial market valuations; Growth accounting; Information technology; Intangible capital; Labour productivity; National income accounting; National Income and Product Accounts (USA); Total factor productivity

JEL Classifications

E21

Economists have long understood that advances in knowledge and technology play a crucial role in economic growth. An important recent contribution to this literature is research on the magnitude and role of intangible capital. As defined by Corrado et al. (2005a, 2006), intangible investment is expenditures by businesses that are intended to boost output in the future but that are not traditional, tangible physical capital; examples include outlays for computer software, research and development, training, brand equity, and improvements in organizational structure and efficiency.

Recent interest in intangible capital was generated by a sense in some quarters that official statistics may not be capturing the full dynamism of the US economy as well as by the resurgence of US productivity growth in the mid-1990s. That resurgence led many researchers, including Oliner and Sichel (2000, 2002), Jorgenson and Stiroh (2000), and Jorgenson et al. (2002), to focus on

the contribution of information technology (IT) to economic growth. And that focus on IT, as well as the run-up in equity valuations that occurred at about the same time, turned researchers' attention to intangible capital. Many analysts observed that firms using IT effectively did more than simply install it; they made sizable collateral investments to revamp their operations in order to exploit the new technologies. For example, Walmart developed a more efficient supply chain, Dell linked demand and production more tightly, Amazon pioneered a new distribution channel, and Google and eBay developed entirely new businesses. In each case, the collateral investments consisted largely of expenditures on intangible inputs. Many observers believe that these intangible investments, as well as intangible investments that may not be tied to IT, are playing an increasingly important role in the economy.

Despite the apparent importance of investments in intangible capital, relatively little is known about these investments. At the firm level, financial accounting provides little information about such expenditures and the return earned by them. Moreover, these outlays are considered a current-period expense, not an investment creating an asset on the firm's balance sheet. Because of this lack of information, Lev (2004) argues that managers may make poor investment decisions and financial markets may incorrectly value firms and therefore may inefficiently allocate capital. At the level of the National Income and Product Accounts (NIPAs) used to measure gross domestic product (GDP) in the United States, historical practice has classified such expenditures as intermediate inputs, and thus they are not counted as investment in GDP. (The inclusion of business software as an investment in the NIPAs is a notable exception to this practice.) Moreover, the GDP accounts, like firm-level financial accounts, provide very little information about most intangible expenditures.

Research has begun to fill this gap with three broad approaches to measuring intangible capital. The first uses financial market valuations to gauge the value of intangible capital, inferring a measure of intangible capital from the gap between the market and book value of firms. As summarized in Hall (2005), such an estimate was quite large around 2000, about equal to the stock of tangible capital. At the firm level, Brynjolfsson and Hitt (2005) regress market value on capital and labour inputs as well as various proxies for intangible capital. Their work highlights the link between intangible investments and investments in computers, and suggests that intangible investments may exceed tangible investments in computers by as much as a factor of ten. Considerable controversy has surrounded estimates of intangible capital that are derived from financial market valuations.

The second broad category of research relies on other performance measures (such as productivity or earnings) to gauge the magnitude of intangible capital; for examples, see McGrattan and Prescott (2005), Cummins (2005), and Lev and Radhakrishnan (2005). Lev (2004) summarizes a methodology for estimating the value of intangibles at the level of individual firms, starting from earnings. This literature also finds a large role for intangibles.

The third broad category of research uses expenditure data to develop measures of intangible capital. Nakamura (1999, 2001, 2003) was the first to develop expenditure measures. Corrado et al. (2005a) expanded on Nakamura's work and more tightly integrated estimates of intangible investment with the NIPAs. Marrano and Haskel (2006) applied the methodology of Corrado et al. (2006) to the United Kingdom, and obtained similar results.

Corrado et al. (2006) classify business spending on intangibles into three broad groups: computerized information, innovative property and economic competencies. Computerized information consists mainly of computer software. Innovative property includes scientific R&D and non-scientific R&D such as product development expenditures in financial services and in the entertainment industry. Economic competencies include brand equity (advertising) and firm-specific resources such as training and organizational capital. Corrado, Hulten, and Sichel use a variety of data sources to develop time series of nominal expenditures for each category. These

figures suggest that nominal intangible business investment from 2000 to 2003 averaged $1.2 trillion per year, about $1 trillion of which was not counted as investment in the NIPAs.

This research highlights the magnitude and importance of intangibles but does not quantify their contribution to economic growth. This question is taken up in Corrado et al. (2006), which extends their earlier paper, and embeds intangibles in a conventional growth accounting framework. Specifically, Corrado, Hulten, and Sichel develop time series of the real stock of intangible capital for the United States, using their earlier estimates of investment in intangibles. According to their numbers, the nominal stock of intangible capital was about $3.6 trillion in 2003, about $3.1 trillion of which is not included in official measures. These figures imply that official measures may be understating the stock of business capital by roughly 20 per cent.

Corrado et al. (2006) embed their estimates of intangible capital into a standard growth accounting decomposition and present estimates for the period from 1973 to 2003 for the United States. They compare a decomposition based on data that exclude intangibles to one based on data that include intangible assets. Several important results emerge from this analysis. First, the inclusion of intangibles as investment boosts the estimated growth rate of labour productivity in the non-farm business sector by 10–20 per cent relative to a baseline case that completely ignores intangibles. Second, the contribution of intangibles to economic growth has increased dramatically since 1995, and including intangibles has a considerable effect on the composition of the mid-1990s pickup in labour productivity growth. Third, once intangibles are included, greater use of capital (including both tangible and intangible capital) becomes a more important source of growth. This contrasts with the traditional result (when intangibles are largely excluded), where total factor productivity – the residual after accounting for the contributions from labour and capital – plays a larger role. Finally, the majority of the contribution of intangibles comes from categories of intangibles that have received relatively little attention in the past, such as non-scientific R&D and firm- specific resources. Scientific R&D – perhaps the most studied and most 'traditional' category of intangibles – accounts for only about one-tenth of the contribution of intangibles to labour productivity growth.

Taken together, the research indicates that business investment in intangible capital is quite sizable and has played an important role in the US economy. Moreover, these results indicate that both firm-level and national income accounting practice miss some important features of economic activity. Nevertheless, the quantitative estimates discussed here are clearly provisional, and this area appears to be a fruitful one for further research.

See Also

- ▶ Growth Accounting
- ▶ Intellectual Property

Bibliography

Brynjolfsson, E., and L. Hitt 2005. Remarks. In Corrado, Haltiwanger, and Sichel (2005).

Corrado, C.A., C.R. Hulten, and D.E. Sichel. 2005a. Measuring capital and technology: An expanded framework. In Corrado, Haltiwanger, and Sichel (2005b).

Corrado, C.A., J. Haltiwanger, and D. Sichel, eds. 2005b. *Measuring capital in the new economy.* Chicago: University of Chicago Press.

Corrado, C.A., C.R. Hulten, and D.E. Sichel. 2006. Intangible capital and economic growth. Finance and Economics Discussion Series Paper No. 2006–24. Washington, DC: Federal Reserve Board.

Cummins, J.G. 2005. A new approach to the valuation of intangible capital. In Corrado, Haltiwanger, and Sichel (2005).

Hall, R.E. 2005. Remarks. In Corrado, Haltiwanger, and Sichel (2005).

Jorgenson, D.W., and K.J. Stiroh. 2000. Raising the speed limit: U.S. economic growth in the information age. *Brookings Papers on Economic Activity* 2000(1): 125–211.

Jorgenson, D.W., M.S. Ho, and K.J. Stiroh. 2002. Projecting productivity growth: lessons from the U.S. growth resurgence. *Federal Reserve Bank of Atlanta, Economic Review* 87(3): 1–14.

Lev, B. 2004. Sharpening the intangibles edge. *Harvard Business Review* 82(6): 109–116.

Lev, B., and S. Radhakrishnan. 2005. The valuation of organizational capital. In Corrado, Haltiwanger, and Sichel (2005).

Marrano, M.G., and J. Haskel. 2006. How much does the UK invest in intangible assets? Working Paper No. 578, Department of Economics, Queen Mary College, University of London.
McGrattan, E.R., and E.C. Prescott. 2005. Taxes, regulations, and the value of U.S. and U.K. corporations. *Review of Economic Studies* 72: 767–796.
Nakamura, L. 1999. Intangibles: What put the *new* in the new economy? Federal Reserve Bank of Philadelphia. *Business Review,* July–August, pp. 3–16.
Nakamura, L. 2001. What is the US gross investment in intangibles? (At least) one trillion dollars a year! Working Paper No. 01–15, Federal Reserve Bank of Philadelphia.
Nakamura, L. 2003. The rise in gross private investment in intangible assets since 1978. Mimeo, Federal Reserve Bank of Philadelphia.
Oliner, S.D., and D.E. Sichel. 2000. The resurgence of growth in the late 1990s: Is information technology the story? *Journal of Economic Perspectives* 14(4): 3–22.
Oliner, S.D., and D.E. Sichel. 2002. Information technology and productivity: Where are we now and where are we going? *Federal Reserve Bank of Atlanta Economic Review* 87(3): 15–44.

Integer Programming

Egon Balas

Integer programming is the youngest branch of mathematical programming: its development started in the late 1950s. A (linear or nonlinear) *integer program* is a linear or nonlinear program whose variables are constrained to be integer. We will consider here only the linear case, although there exist extensions of the techniques to be discussed to nonlinear integer programming.

The integer programming problem can be stated as

$$(\mathrm{p}) \times \min\{cx | Az \geq b, x \geq 0, x_j \text{integer}, j \in N_1 \subseteq N\},$$

where A is a given $m \times n$ matrix, c and b are given vectors of conformable dimensions, $N = \{1,\ldots, n\}$, and x is a variable n-vector. (P) is called a *pure integer program* if $N_1 = N$, a *mixed integer program* if $\varphi \neq N_i \neq N$. Integer programming is sometimes called *discrete optimization*.

Modelling Potential

Integer programming is the most immediate and frequently needed extension of linear programming. Integrality constraints arise naturally whenever fractional values for the decision variables do not make sense. A case in point is the *fixed charge problem*, in which one wants to minimize a function of the form $\Sigma_j c(x_i)$, with

$$c(x_j) = \begin{cases} f_j + c_j x_j & \text{if } x_j > 0 \\ 0 & \text{if } x_i = 0 \end{cases}$$

subject to linear constraints. Such a problem can be restated as an integer program whenever x is bounded by setting

$$c(x_i) = c_i x_i + f_i y_i$$
$$x_i \leq U_i y_i, \quad y_i = 0 \text{ or } 1$$

where U_i is an upper bound of x_i.

By far the most important special case of integer programming is the *0–1 programming problem*, in which the integer-constrained variables are restricted to 0 or 1. This is so because a host of frequently occurring nonlinearities, like logical alternatives, implications, precedence relations, etc., or combinations thereof, can be formulated via 0–1 variables. For example, a condition like

$$x > 0 \Rightarrow (f(x) \leq a \vee f(x) \geq b),$$

where a and b are positive scalars, x is a non-negative variable with a known upper bound M, $f(x)$ is a function whose value is bounded from above by $U > b$ and from below by $L < a$, while the symbol '$\vee$' means disjunction (logical 'or'), can be stated as

$$x \leq M(1 - \delta_1)$$
$$F(x) \leq a + (U - a)\delta_1 + (U - a)\delta_2$$
$$F(x) \geq b + (L - a)\delta_1 + (L - b)(1 - \delta_2)$$
$$\delta_1 \delta_2 = 0 \text{ or } 1$$

A linear program with *logical* conditions (conjunctions, disjunctions and implications) involving inequalities is called a *disjunctive program*, since it is the presence of disjunctions that

makes these problems nonconvex. Bounded disjunctive programs can be stated as 0–1 programs and vice versa, but the disjunctive programming formulation has led to new methods.

Nonconvex optimization problems like bimatrix games, separable programs involving piecewise linear nonconvex functions, the general (nonconvex) quadratic programming problem, the linear complementarity problem and many others can be stated as disjunctive or 0–1 programming problems.

A host of interesting combinatorial problems can be formulated as 0–1 programming problems defined on a graph. The joint study of these problems by mathematical programmers and graph theorists has led to the recent development of a burgeoning area of research known as *combinatorial optimization*. Some typical problems studied in this area are: edge matching and covering, vertex packing and covering, clique covering, vertex colouring; set packing, partitioning and covering; Euler tours; Hamiltonian cycles (travelling salesman problem).

Applications of integer programming abound in all spheres of decision making. Some typical real-world problem areas where integer programming is particularly useful as a modelling tool include: facility (plant, warehouse, hospital, fire station) location; scheduling (of personnel, machines, projects); routing (of trucks, tankers, airplanes); design of communication (road, pipeline, telephone) networks; capital budgeting; project selection; analysis of capital development alternatives.

Solution Methods

Integer programs are notoriously hard to solve. Unlike linear programs, which are always solvable in a number of steps bounded by a polynomial function of the length of the data input, integer programs often require a number of steps that grows exponentially with problem size. However, sometimes an integer program (P) can be solved as a linear program; i.e., solving the linear program (L) obtained by removing the integrality constraints from (P), one obtains an integer solution. In particular, this is the case when all basic solutions of (L) are integer. For an arbitrary integer vector b, the constraint set $Ax \leq b$, $x \geq 0$ is known (Hoffman and Kruskal 1958) to have only integer basic solutions if and only if the matrix A is totally unimodular (i.e. all nonsingular submatrices of A have a determinant of 1 or -1).

The best known instances of total unimodularity are the vertex-edge incidence matrices of directed graphs and undirected bipartite graphs. As a consequence, shortest path and network flow problems on arbitrary directed graphs, edge matching (or covering) and vertex packing (or covering) problems on bipartite graphs, as well as other integer programs whose constraint set is defined by the incidence matrix of a directed graph or an undirected bipartite graph, with arbitrary integer right-hand side, are in fact linear programs.

Apart from this important but very special class of problems, and a few other special classes, the difficulty in solving integer programs lies in the nonconvexity of the feasible set, which makes it impossible to establish global optimality from local conditions. The two main approaches to solving integer programs try to circumvent this difficulty in two different ways.

The first approach, which in the current state of the art is the standard way of solving general integer programs, is *enumerative* (branch and bound, implicit enumeration). It partitions the feasible set into successively smaller subsets, calculates bounds on the objective function value over each subset, and uses these bounds to discard certain subsets from further consideration. The procedure ends when each subset has either produced a feasible solution, or was shown to contain no better solution than the one already in hand. The best solution found during the procedure is a global optimum. Two early prototypes of this approach are due to Land and Doig (1960) and Balas (1965).

The second approach, known as the cutting plane method, is a *convexification* procedure: it approximates the convex hull of the set F of feasible integer points, by a sequence of inequalities that cut off (hence the term ‘cutting planes’) parts of the linear programming polyhedron, without

removing any point of *F*. When sufficient inequalities have been generated to cut off every fractional point better than the integer optimum, then the latter is found as an optimal solution to the linear program (L) amended with the cutting planes. The first finitely convergent procedure of this type is due to Gomory (1958).

Depending on the type of techniques used to describe the convex hull of *F* and generate cutting planes, one can distinguish three main directions in this area. The first one uses algebraic methods, like modular arithmetic and group theory. Its key concept is that of subadditive functions. It is sometimes called the algebraic or group theoretic approach. The second one uses convexity, polarity and propositional calculus. Its main thrust comes from looking at the 0–1 programming problem as a disjunctive program. It is known as the convex analysis/disjunctive programming approach. Finally, the third direction applies to combinatorial programming problems, and it combines graph theory and matroid theory with mathematical programming. It is sometimes called polyhedral combinatorics.

Besides these two basic approaches to integer programming (enumerative and convexifying), two further procedures need to be mentioned, that do not belong to either category, but can rather be viewed as complementary to one or the other. Both procedures essentially *decompose* (P), one of them by partitioning the variables, the other one by partitioning the constraints. The first one, due to Benders (1962) eliminates the continuous variables of a mixed integer program (P) by projecting the feasible set *F* into the subspace of the integer-constrained variables. The second one, known as Lagrangean relaxation, removes some of the constraints of (P) by assigning multipliers to them and taking them into the objective function.

Each of the approaches outlined here aims at solving (P) exactly. However, since finding an optimal solution tends to be expensive beyond a certain problem size, approximation methods or *heuristics* play an increasingly important role in this area.

At present all commercially available integer programming codes are of the branch and bound type. While they can *sometimes* solve problems with hundreds of integer and thousands of continuous variables, they cannot be *guaranteed* to find optimal solutions in a reasonable amount of time to problems with more than 30–40 variables. On the other hand, they usually find feasible solutions of acceptable quality to much larger problems.

A considerable number of specialized branch and bound/implicit enumeration algorithms have been implemented by operations research groups in universities or industrial companies. They usually contain other features besides enumeration, like cutting planes and/or Lagrangean relaxation. Some of these codes can solve general (unstructured) 0–1 programs with up to 80–100 integer variables, and structured problems with up to several hundred (assembly line balancing, multiple choice, facility location), a few thousand (sparse set covering or set partitioning, generalized assignment), or several thousand (knapsack, travelling salesman) 0–1 variables.

At the current state of the art, while many real-world problems amenable to an integer programming formulation fit within the stated limits and are solvable in useful time, others substantially exceed those limits. Furthermore, some important and frequently occurring realworld problems, like job shop scheduling and others, lead to integer programming models that are almost always beyond the limits of what is currently solvable. Hence the great importance of approximation methods for such problems.

See Also

- ▶ Combinatorics
- ▶ Indivisibilities
- ▶ Linear Programming

Bibliography

Balas, E. 1965. An additive algorithm for solving linear programs with zero–one variables. *Operations Research* 13: 517–546.

Benders, J.F. 1962. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik* 4: 238–252.

Gomory, R.E. 1958. Outline of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Society* 64: 275–278.

Hoffman, A.J., and J.B. Kruskal. 1958. Integral boundary points of convex polyhedra. In *Linear inequalities and related systems*, ed. H.W. Kuhn and A.W. Tucker, 223–246. Princeton: Princeton University Press.

Land, A.H., and A.G. Doig. 1960. An automatic method for solving discrete programming problems. *Econometrica* 28: 497–520.

Integrability of Demand

Donald W. Katzner

Abstract

Integrability of demand arguments start with consumer demand functions having properties that would be implied by constrained utility maximization were they generated from that source. Using a process of mathematical integration, the arguments then proceed to demonstrate the existence of utility functions from which those demand functions could be derived.

Keywords

Demand function; Integrability of demand; Marginal rate of substitution; Revealed preference; Slutsky substitution functions; Utility function

JEL Classifications

D11

The lines of reasoning linking individual (ordinal) utility functions (or preference orderings) to individual demand functions run in both directions. Progressions from the former to the latter often begin with assumptions about the characteristics of a consumer's utility function and the requirement that he or she always chooses so as to maximize utility subject to a budget constraint, and then go on to derive the demand functions and the properties of those demand functions that logically ensue from such premises. Depending on context, certain of the properties of the demand functions so derived are expressed in differential terms (that is, symmetry and negative definiteness of matrices of Slutsky substitution functions where the latter are defined) or in discrete revealed preference form (for example, weak and strong axioms of revealed preference). The reverse course takes the individual's demand functions and their properties as given and determines the existence of a utility function from which, upon constrained maximization, the original demand functions could have been generated. In this second case, when the starting point includes the differential rather than revealed preference properties of demand, the argument often involves (in part) the integration of a system of one or more differential equations. Hence the name 'integrability of demand' affixed to it.

There are several ways to structure an integrability of demand argument. Perhaps the most straightforward approach (the only one considered in detail here) is simply to backtrack over the path that yields demand functions from utility functions via the theorem of Lagrange on maximization subject to constraint. That path may be summarized as follows. Begin with a utility function $\mu = u(x)$ defined over the commodity space $\{x : x \geq 0\}$, where $x = (x_1,\ldots,x_I)$ is a vector of quantities of commodities x_i and $x \geq 0$ means $x_i \geq 0$ for every $i = 1,\ldots, I$. Let $u_i(x)$ be the partial derivative of the utility function u with respect to its i^{th} argument. For each vector $(p, m) > 0$, where $p = (p_1,\ldots, p_I)$, p_i is the price of good i, and m is a scalar denoting the consumer's income, vectors $x > 0$ that maximize $u(x)$ subject to the budget constraint $\sum_{i=1}^{I} p_i x_i = m$ are, according to Lagrange's theorem, characterized by

$$\frac{p_i}{p_I} = \frac{u_i(x)}{u_I(x)}, \qquad i = 1,\ldots,I-1, \tag{1}$$

$$\frac{m}{p_I} = x_I + \sum_{i-1}^{I-1} \frac{u_i(x)}{u_I(x)}\, x_i. \tag{2}$$

Equations (1) state that, at a constrained maximum, the marginal rates of substitution or the negatives of the partial derivatives of indifference functions equal the price ratios, and Eq. (2) is a form of the budget constraint. Equations (1) and (2) together may be thought of as a system of inverse functions which are solved to secure demands x_i as functions h^i, of prices and income:

$$x_i = h^i\left(\frac{p_1}{p_I},\ldots,\frac{p_{I-1}}{p_I},\frac{m}{p_I}\right), \quad i = 1,\ldots,I. \tag{3}$$

Evidently, (3) may be written in the equivalent form

$$x_i = H^i(p,m), \qquad i = 1,\ldots,I,$$

where $H^i(p, m) = h^i(p_1/p_I,\ldots,p_{I-1}/p_I, m/p_I)$ and H^i is homogeneous of degree zero. Of course, sufficient properties have to be imposed on u so as to ensure the existence of a usually unique constrained maximizing x for each $(p, m) > 0$, and these properties, in turn, imply the well-known characteristics of the h^i or the H^i.

Consider now an integrability of demand argument that reverses the above steps. Start with the demand functions (3) having all of the properties that would be implied by constrained utility maximization were they determined as previously described. The aim is to show the existence of a utility function generator of these demand functions. Clearly, for this latter utility function to generate the h^i, it must exhibit properties such as those stated above that yield unique constrained maxima. Backtracking from (3), solve for price–price and income–price ratios as functions, g^i, of x:

$$\frac{p_1}{p_I} = g^i(x), \qquad i = 1,\ldots I-1, \tag{4}$$

$$\frac{m}{p_I}\, g^I(x). \tag{5}$$

If the h^i are to be derivable from constrained utility maximization, then the g^i of (4) should indicate the negatives of the partial derivatives of appropriate indifference functions and g^I should be related to the budget constraint; that is, Eq. (4) should correspond to Eq. (1), and Eq. (5) to Eq. (2). But to say that the g^i are the negatives of the partial derivatives of indifference functions means that

$$\frac{\partial x_1}{\partial x_i} = -g^i(x), \quad i = 1,\ldots,I-1. \tag{6}$$

Thus, at every $x > 0$, the 'slopes' of the indifference surface through x in the direction of each of the coordinate axes are given by the g^i, for $i = 1,\ldots,I-1$. Integrating the differential equation system (6) 'fits' all of these slopes for each surface together to form an indifference map from which a utility function is deduced (It is possible to integrate alternative, though related, systems of differential equations which yield the utility function directly). It should be noted, however, that the mathematics employed in this integration process usually shows only that such an indifference map, and hence a utility function, exists and typically does so without providing the means to specify the exact forms it will take. Lastly, the appropriate general characteristics of this utility function and the fact that its constrained maximization produces the original demand functions (3) are established.

Naturally, the properties of the demand functions h^i are crucial for such an integrability of demand argument to hold up. Among other things, these properties must permit the inversion of the h^i into the g^i and must ensure that the integration step can be carried out. Invertibility means that the h^i specify a 1–1 correspondence between values of the vectors $x = (x_1,\ldots,x_I)$ and $(p_1/p_I,\ldots,p_{I-1}/p_I, m/p_I)$. For the integration of (6) it is necessary that the g^i be continuous and, when $I > 2$, that a certain 'integrability' condition be satisfied. This guarantees that at least one indifference surface passes through every $x > 0$. To make certain that no more than one indifference surface passes through each x, a Lipschitz condition has to be in force. It turns out that the g^i are continuous as long as the h^i are continuous, that the integrability condition is equivalent to the existence and symmetry, for all $(p, m) > 0$, of the matrix of Slutsky substitution functions, and

that the Lipschitz condition is implied if certain partial derivatives of the g^i are bounded. All of these properties of demand functions except the last two are derivable from the constrained maximization of utility functions that are twice continuously differentiable, increasing, strictly quasiconcave, and whose indifference surfaces do not touch the boundaries of the commodity space (Although such utility-function characteristics imply symmetry of the matrix of Slutsky substitution functions, they do not guarantee that those functions, and hence the matrix, will be defined everywhere). Even so, the properties of demand functions obtained from such utility functions are still 'roughly' sufficient to support the integrability of demand argument outlined above.

Problems arise when the properties of demand functions are derived from utility functions with modified characteristics. For example, the previously mentioned 1–1 correspondence may not appear in the h^i and hence invertibility from the h^i to the g^i may break down. In such a situation it is possible to restructure the integrability of demand argument to avoid the invertibility issue at the level of the h^i altogether. Since it turns out that the demand functions $H^i(p, m)$ may also be viewed as partial derivatives with respect to p_I of the expenditure or income compensation function (obtained in the progression from utility to demand by minimizing expenditure for a given level of utility)

$$m = E(p, \mu),$$

where μ varies over all utility levels and p ranges over all vectors $p > 0$, this is accomplished by integrating the system

$$\frac{\partial m}{\partial pi} = H^i(p, m), \quad i = 1, \ldots, I, \tag{7}$$

and converting the resulting expenditure function into a utility function. Once again the appropriate characteristics of the derived utility function have to be established, constrained maximization of it has to produce the given H^i, and enough properties of the H^i need to be present to sustain the argument.

Antonelli (1886) is usually credited with introducing economists to the integrability of demand argument. He began with the functions g^i and obtained a utility generator by integrating a system of differential equations related to (6). Many years later in a mathematical appendix, Samuelson (1950) inverted the h^i and then secured an indifference map by integrating another differential equation related to (6). In between, Antonelli's work seems to have been almost forgotten. Fisher (1892) independently 'rediscovered' the integrability problem in his doctoral dissertation, and various aspects of it were taken up subsequently by Pareto (1906a, b), Volterra (1906), Allen (1932), Georgescu-Roegen (1936), Wold (1943, 1944), and others. It is interesting that Volterra's contribution was to point out that, in Pareto's initial (1906a) discussion of integrability for the case of more than two goods (that is, in the first edition of his *Manuale*), the integrability condition had been conspicuously omitted. More detailed history is given by Samuelson (1950) and Chipman et al. (1971, intro. to Part II). Hurwicz and Uzawa (1971) were the first to structure an integrability of demand argument based on the integration of (7).

Bibliography

Allen, R. 1932. The foundations of a mathematical theory of exchange. *Economica* 12: 197–226.

Antonelli, G. 1886. On the mathematical theory of political economy. In *Preferences, utility, and demand,* ed. J. Chipman et al. New York: Harcourt Brace, Jovanovich, 1971.

Chipman, J., L. Hurwicz, M. Richter, and H. Sonnenschein, eds. 1971. *Preferences, utility, and demand*. New York: Harcourt Brace.

Fisher, I. 1892. Mathematical investigations in the theory of value and prices. *Transactions of the Connecticut Academy of Arts and Sciences* 9(July): 1–124. Reprinted, New York: Augustus M. Kelley, 1965.

Georgescu–Roegen, N. 1936. The pure theory of consumer's behavior. *Quarterly Journal of Economics* 50: 545–593.

Hurwicz, L., and H. Uzawa. 1971. On the integrability of demand functions. In *Preferences, utility, and demand*, ed. J. Chipman et al. New York: Harcourt Brace, Jovanovich.

Katzner, D. 1970. *Static demand theory.* New York: Macmillan.

Pareto, V. 1906a. *Manuale di economia politica*. Milan: Societa Editrice Libraria.
Pareto, V. 1906b. Ophelimity in nonclosed cycles. In *Preferences, utility, and demand,* ed. J. Chipman et al. New York: Harcourt Brace, Jovanovich, 1971.
Samuelson, P. 1950. The problem of integrability in utility theory. *Economica* 17: 355–385.
Volterra, V. 1906. Mathematical economics and Professor Pareto's new manual. In *Preferences, utility, and demand,* ed. J. Chipman et al. New York: Harcourt Brace, Jovanovich, 1971.
Wold, H. 1943. A synthesis of pure demand analysis I, II. *Skandinavisk Aktaurietidskrift* 26 (85–118): 221–263.
Wold, H. 1944. A synthesis of pure demand analysis III. *Skandinavisk Aktaurietidskrift* 27: 69–120.

Intellectual Property

Michele Boldrin and David K. Levine

Abstract

Intellectual property refers to patents, copyrights, trademarks and other forms of ownership of ideas. It results in monopoly power that has significant consequences for discouraging as well as encouraging innovation and growth. The discouragement effect is especially important when ideas are used as building blocks for other ideas. The economics literature has examined the need for intellectual property; optimal systems of intellectual property; the optimal duration of intellectual property; how innovation takes place in the absence of intellectual property; and the rent-seeking behaviour induced by intellectual property.

Keywords

Arrow, K.; Copyright; Dynamic vs static efficiency; First-mover advantage; Innovation, competitive; Monopoly vs incentive to innovate; No-compete contract clauses; Non-disclosure agreements; Patent law; Patent races; Patents; Plant, A.; Rent seeking; Schumpeter, J.; Shrink-wrap agreements; Stigler, G.; Trademarks; Transaction costs

JEL Classifications

O3

Intellectual property refers to patents, copyrights, trademarks and other forms of ownership of ideas. While property and ownership are not controversial topics among economists, patents and copyrights have long been. By contrast, trademarks – serving merely to identify individuals and businesses – are not controversial. The economic analysis of patents and copyrights applies also to a variety of private contractual arrangements that are used to enforce 'intellectual property' such as non-disclosure agreements, no-compete contract clauses, and software shrink-wrap agreements.

The controversy surrounding patents and copyright has both theoretical and policy relevance. The theoretical relevance arises because models of economic growth, trade, and industrial regulation all put innovative activity at their core. Two fundamental views of innovation have been advanced. In the first, which can be traced back to Arrow (1962) and has recently been developed by Romer (1986, 1990), it is abstract ideas that matter. These are produced subject to fixed costs that cannot be recouped under competition because, once discovered, ideas are non-rivalrous and infinitely reproducible at constant marginal cost, which is often treated as equal to zero. In the second, traces of which are found in Schumpeter (1911), Plant (1934) and Stigler (1956), and is formalized by Boldrin and Levine (2003), it is the concrete embodiment of copies of ideas that has economic value. Embodied ideas are initially characterized by indivisibility, but their reproduction is limited by capacity constraints; the latter generate competitive rents that may cover the indivisibility cost, hence ideas can be produced and traded under competition.

From both perspectives, 'intellectual property' is a grant of monopoly power over the right to make copies of ideas, and not simply the extension of 'normal' property rights to the realm of ideas. The first view argues in favour of intellectual property not because it is property and serves to protect the value of individual investment, but

rather because monopoly over ideas can be a good thing. This argument has recently been linked, by such authors as Aghion and Howitt (1992) or Grossman and Helpman (1991), to a Schumpeterian theme (Schumpeter 1942). It posits a trade-off between 'static efficiency', which requires competition, and 'dynamic efficiency', which can be achieved only through technological progress driven, in turn, by the desire to acquire a monopoly. In this view it is monopoly power that drives the innovative process.

On the policy side, intellectual property has become controversial largely because of three developments. The first is the high price and restrictive policies of pharmaceutical companies, for example with regard to AIDS drugs. Second is the damaging impact of intellectual property on the growth perspectives of the less developed countries, especially when they are denied free trade with the more developed ones unless they adopt strict standards for intellectual property. Third is the impact of the internet on the 'piracy' of music, books and movies.

On one side of the policy debate stand those who benefit from existing monopolies and are eager to protect their way of doing business, arguing typically that their 'property' should be protected from 'theft'. On the other side is a broad array of people who resent having the free use of their copies of ideas restricted by creators.

The central issue is whether the monopoly power achieved through copyrights and patents is truly necessary, in the words of the US Constitution, 'To promote the progress of science and useful arts' or whether it in fact hinders progress and innovation.

Optimal Systems of Intellectual Property

A key question is what an optimal system of intellectual property looks like. Most research has focused on patents, and much attention has been devoted to the issue of the breadth versus the duration of patents – that is, whether long but narrow patents are preferable to short but broad ones. The seminal paper on the subject is that of Gilbert and Shapiro (1990), which models breadth as a price ceiling limiting the patent holder's ability to price at the full monopoly price. In the Gilbert–Shapiro setup, the conclusion is that optimal patents should be as long and narrow as possible. Subsequent authors have contested this conclusion – Gallini (1992), in particular, argues that the model of 'breadth' as a price ceiling does not reflect what 'breadth' is likely to mean in practice, and that a more reasonable model of breadth leads to the opposite conclusion. Subsequent work by Gallini and Scotchmer (2001) concludes that optimal patent protection should probably be broad but short.

Recent research, for example by Hopenhayn and Mitchell (2001), has examined the possibility of providing not a single breadth-cum-duration for all patents, but rather allowing patent applicants to select from a menu of alternatives: some might choose broad but short, others narrow but long. They show that a properly calibrated system of this type can be superior to a one-size-fits-all system.

Many of these models recognize that innovators will earn something even without patent protection, and it is generally true in these models that if enough rents are earned without patents the optimal system is to have no patents at all. From the perspective of designing a system, this answer – no patents – is not interesting, and for understandable reasons this case tends to be underplayed. However, while how not to design a patent system may be less intellectually challenging than how to design one, the case in which adequate rents are earned without patents may well be empirically more relevant.

Optimal Duration of Intellectual Property

Quite apart from the details of system design, the question arises of how much protection an IP system should provide. Why, for example, should monopoly be limited rather than unrestricted? Answering this question requires trading off the monopoly power created by intellectual property against the incentive to innovate. This issue was first studied in the context of

copyrights by Ian and Waldman (1984). They examine a model in which 'innovation' has the dimension of higher product quality. However, they assume away the harmful effects of monopoly power by assuming that demand is completely elastic up to an upper bound. Not surprisingly, in this setting stronger intellectual property is unambiguously good. Beginning with Liebowitz (1985), Stan Liebowitz has also extensively studied copyrights; focusing largely on a single creation, he argues that the indirect appropriability of competitive rents is generally an inadequate incentive to create.

More recently, Grossman and Lai (2004) and Boldrin and Levine (2005a) have examined a general equilibrium setting in which ideas of different quality are produced. Both papers show that optimal protection is generally limited rather than unlimited; and, because markets are growing over time, they consider the consequences of expanded markets for optimal protection. Boldrin and Levine show that optimal protection always declines when the market is large enough. They also give an elasticity condition under which protection should always decline with market size and, based on examination of existing data, conclude that this condition is likely to be satisfied in practice. This empirical analysis builds heavily on an empirical literature, stemming from Pakes (1986), that tries to estimate the distribution of patent values by examining such things as patent renewal rates.

Patent Races

Some time ago, theoretical work focused on patent races, in which firms over-invest in R&D in an effort to obtain a valuable patent before a rival. Fudenberg et al. (1983) and Harris and Vickers (1985) are two of the earliest papers on these lines. Contrary to the traditional problem of too little innovation, this line of research suggests that the desire to acquire the monopoly power that patents confer may encourage too much expenditure in wasteful R&D. These models seem to have fallen out of favour in recent years, perhaps because there is little empirical evidence that patent races are quantitatively important in determining the pace of actual technological innovation. The fact that legal battles over patents have become a persistent feature of contemporary business strategy may well restore currency to a modified version of these models.

Ideas as Building Blocks

One critical element of innovation – and artistic creation as well – is that new ideas are generally built upon existing ideas. That this is true for patentable innovations is fairly obvious (in the realm of copyright see Lessig 2004, and Vaidhyanathan 2003). Scotchmer (1991) points out that strong patent protection can have the dual effect of increasing the return to innovation and at the same time increasing the cost of acquiring the rights needed to innovate. This point is developed further in Boldrin and Levine, who show that under certain conditions a patent system may serve only to discourage innovation (2003), and that, when the innovator is better informed about the value of a new idea than the holders of rights to previous ideas, a patent system serves strictly to discourage innovation (2005b). Intuitively, all the additional profit from the new innovation is absorbed by the existing rights holders; if there are many of them, there is a public goods problem, with each 'little monopolist' setting a price that is too high because much of the cost of decreased likelihood of innovation is borne by the other 'little monopolists'. This type of holdup problem is not dissimilar to the problem pointed out by Chari and Jones (2000) in the context of externalities more broadly; interestingly, in this case externalities are created by the existence of intellectual property, and would be altogether absent without it.

The practical impact of intellectual property in a setting where the use of existing ideas is important is well documented by Bessen and Hunt (2003), who examine the software industry in the United States during the era of personal computers; they find that intellectual property has been antithetical to innovation in this industry.

The role of transactions costs that arise when it is necessary to acquire many rights in order to innovate is underlined by David Friedman's (1994) striking hypothetical example of what

would happen if every word in the English language was copyrighted, so that any writer had to pay for each use of every word.

Competitive Innovation

Since there is a well-documented downside to intellectual property, it is important to understand how markets might function in its absence. Arnold Plant and George Stigler, among others, provide important examples of innovation and creation taking place without the benefit of monopoly. Plant (1934, p. 173) writes that, although in the 19th century English authors could not copyright their works in the United States,

> . . . American publishers found it profitable to make arrangements with English authors . . . English authors sometimes received more from the sale of their books by American publishers, where they had no copyright, than from their royalties in [England].

Similarly, Stigler (1956, p. 274) argues that monopoly is completely unnecessary to provide incentives for innovation.

> There can be rewards – and great ones – to the successful competitive innovator. For example, the mail-order business . . . The innovators . . . were Aaron Montgomery Ward, who opened the first general merchandise establishment in 1872, and Richard Sears . . . Sears soon lifted his company to a dominant position by his magnificent merchandising talents, and he obtained a modest fortune, and his partner Rosenwald an immodest one. At no time were there any conventional monopolistic practices, and at all times there were rivals within the industry and other industries making near-perfect substitutes . . .

In more recent times, Liebowitz (1985), Boldrin and Levine (2003), Quah (2002), Legros (2005), and Hellwig and Irmen (2001) have all examined the competitive rents that accrue to innovators due to 'limited capacity' – the fact that in a competitive market the owners of a fixed factor (first copy of an idea) are the recipients of all downstream rents originating from it, and that an infinite number of copies cannot be made instantaneously. The conclusion is that innovation will take place even without intellectual property – as it often has in the past (see for example the cases mentioned by Moser 2002). While some of this work shows that there may be too little innovation under competition due to the indivisible nature of the initial copy of ideas, it also suggests that the appropriate remedy is unlikely to be a government-granted monopoly.

In modern times, evidence that patents are unnecessary to provide the adequate incentive to innovate can be found in the widespread cross-licensing agreements found in chip manufacturing. The evidence is discussed by Shapiro (2001), who argues that the sharing of information between chip firms is much more important to them than any short-term advantage gained through a patent, and the primary function of patenting in this industry is to block entry by potential rivals.

First-Mover Advantage

Regardless of the presence of competitive rents, an innovator is likely to have a substantial advantage by being first to the market. This is largely what Plant and Stigler had in mind. The important impact of first-mover advantage in the market for new types of financial securities prior to the advent of patents in that industry has been ably documented by Tofuno (1989), and the theory explained carefully, together with further evidence, by Herrera and Schroth (2002, 2003).

Besides the temporary monopoly that results from being first, there are less obvious advantages. Hirshleifer (1971) first, and Anton and Yao (1994) subsequently, show how advance knowledge of an innovation can give an edge in asset markets. This can be illustrated through the example of the 'Segway' scooter – much publicized when it was introduced as a revolution in transportation. Suppose for the moment that these claims were true: how could the inventor have profited from this information without – as he did – surrounding himself with a thicket of patents? The Hirshleifer scheme would see him selling short automobile stocks, which would drop through the floor as soon as he announced his discovery. The Anton–Yao scheme would have the inventor sell the idea to, say, Ford in exchange for a share of the profits. Since he would share the

profits, he would then have no incentive to try to sell the idea to other automobile companies, and so Ford would be happy to pay for the resulting monopoly. If it simply took the idea without paying, Ford would lose the monopoly when the inventor told the other companies how to build Segways. Along similar lines, Baccara and Razin (2004) have developed a bargaining model for the case when the inventor must share the idea with others to implement it, and the latter can 'run away' after the idea is revealed. Even in the absence of any intellectual property, as the idea is revealed to more and more people and market power dilutes, the 'threat of competition' is enough to make the collaborator comply and to guarantee the innovator a substantial (larger than one-third) share of the surplus.

Rent Seeking

The most significant downside of government grants of monopoly is the rent seeking they trigger. For example, although they have a generally favorable view of patents, historical research by Lamoreaux and Sokoloff (2001) shows that tightening of patent law resulted in a large upswing in innovation – presumably because it eliminated the nuisance of 'submarine' and other patents designed to appropriate value from the true innovators.

Outside the direct line of those whose existing way of business is threatened by innovation, enormous concern has been expressed at the consequences of rent seeking for the limitations it imposes on personal liberty and the threat it poses to economic progress. For example, the efforts of large media giants to 'protect' their 'intellectual property' through government-mandated hardware installed in computers poses a significant threat to innovation in the much larger IT industry.

Throughout history governments with little ability to monitor transactions and collect tax revenue have often fallen back on grants of monopoly to private individuals. Current patent and copyright systems seem to be remnants from this era, and many economists wonder if it is not time to replace them with more efficient modern systems of graduated incentives such as tax subsidies.

See Also

▶ Patents

Bibliography

Aghion, P., and P. Howitt. 1992. A model of growth through creative destruction. *Econometrica* 60: 323–351.

Anton, J., and D. Yao. 1994. Expropriation and inventions: Appropriable rents in the absence of property rights. *American Economic Review* 84: 190–209.

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity*, ed. R. Nelson. Princeton: Princeton University Press.

Baccara, M., and R. Razin. 2004. From thought to practice: Appropriation and endogenous market structure with imperfect intellectual property rights. Discussion Paper No. 4419. London: CEPR.

Bessen, J., and R. Hunt. 2003. An empirical look at software patents. Working paper No. 2003–17. Philadelphia: Federal Reserve Bank of Philadelphia.

Boldrin, M., and D. Levine. 2003. Perfectly competitive innovation. Mimeo. University of Minnesota and UCLA. http://www.econ.umn.edu/~mboldrin/Papers/pci39.pdf. Accessed 14 July 2005.

Boldrin, M., and D. Levine. 2005a. IP and market size. Mimeo. University of Minnesota and UCLA. http://www.dklevine.com/papers/scale22.pdf. Accessed 14 July 2005.

Boldrin, M., and D. Levine. 2005b. The economics of ideas and intellectual property. *Proceedings of the National Academy of Sciences* 102: 1252–1256.

Chari, V., and L. Jones. 2000. A reconsideration of the problem of social cost: Free riders and monopolists. *Economic Theory* 16: 1–22.

Friedman, D. 1994. Standards as intellectual property: An economic approach. *University of Dayton Law Review* 19: 1109–1129.

Fudenberg, D., R. Gilbert, J. Stiglitz, and J. Tirole. 1983. Preemption, leapfrogging and competition in patent races. *European Economic Review* 22: 3–31.

Gallini, N. 1992. Patent policy and costly imitation. *RAND Journal of Economics* 23: 52–63.

Gallini, N., and S. Scotchmer. 2001. Intellectual property: When is it the best incentive system? Economics Working Papers E01–303. Berkeley: University of California.

Gilbert, R., and C. Shapiro. 1990. Optimal patent length and breadth. *RAND Journal of Economics* 21: 106–112.

Grossman, G., and E. Helpman. 1991. Quality ladders in the theory of growth. *Review of Economic Studies* 58: 43–61.

Grossman, G., and E.L.-C. Lai. 2004. International protection of intellectual property. *American Economic Review* 94: 1635–1653.

I

Harris, C., and J. Vickers. 1985. Patent races and the persistence of monopoly. *Journal of Industrial Economics* 33: 461–481.
Hellwig, M., and A. Irmen. 2001. Endogenous technical change in a competitive economy. *Journal of Economic Theory* 101: 1–39.
Herrera, H., and E. Schroth. 2002. Profitable innovation without patent protection: The case of derivatives. Mimeo. Mexico, D.F.: Department of Economics, ITAM. http://ciep.itam.mx/~helios/deriv.pdf. Accessed 14 July 2005.
Herrera, H., and E. Schroth. 2003. Developer's expertise and the dynamics of financial innovation: Theory and evidence. Mimeo. Mexico, D.F.: Department of Economics, ITAM. http://ciep.itam.mx/~helios/express.pdf. Accessed 14 July 2005.
Hirshleifer, J. 1971. The private and social value of information and the reward to inventive activity. *American Economic Review* 61: 561–574.
Hopenhayn, H., and M. Mitchell. 2001. Innovation variety and patent breadth. *RAND Journal of Economics* 32: 152–166.
Ian, E., and M. Waldman. 1984. The effects of increased copyright protection: An analytic approach. *Journal of Political Economy* 92: 236–246.
Lamoreaux, N., and K. Sokoloff. 2001. Market trade in patents and the rise of a class of specialized inventors in the nineteenth-century United States. *American Economic Review. Papers and Proceedings* 91: 39–44.
Legros, P. 2005. Art and internet: Blessing the curse? Mimeo. Bruxelles: ECARES, Université Libre de Bruxelles.
Lessig, L. 2004. *Free culture: How big media uses technology and the law to lock down culture and control creativity.* New York: Penguin Press.
Liebowitz, S.J. 1985. Copying and indirect appropriability: Photocopying of journals. *Journal of Political Economy* 93: 945–957.
Moser, P. 2002. How do patent laws influence innovation? Evidence from nineteenth century world fairs. Mimeo. Sloan School of Management, MIT.
Pakes, A.S. 1986. Patents as options: Some estimates of the value of holding European patent stocks. *Econometrica* 54: 755–784.
Plant, A. 1934. The economic aspect of copyright in books. *Economica* 1: 167–195.
Quah, D. 2002. 24/7 Competitive innovation. Mimeo. London School of Economics. http://econ.lse.ac.uk/staff/dquah/p/0204-247.pdf. Accessed 4 July 2005.
Romer, P. 1986. Increasing returns and long run growth. *Journal of Political Economy* 94: 1002–1003.
Romer, P. 1990. Endogenous technological change. *Journal of Political Economy* 98: S71–102.
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. Cambridge, MA: MIT Press for the NBER.
Schumpeter, J. 1911. *The theory of economic development*, 1934. Cambridge, MA: Harvard University Press.
Schumpeter, J. 1942. *Capitalism, socialism and democracy*. New York: Harper and Brothers.
Scotchmer, S. 1991. Standing on the shoulders of giants: Cumulative research and the patent law. *Journal of Economic Perspectives* 5: 29–41.
Stigler, G. 1956. Industrial organization and economic progress. In *The state of the social sciences*, ed. L. White. Chicago: University of Chicago Press.
Tofuno, P. 1989. First mover advantages in financial innovation. *Journal of Financial Economics* 3: 350–370.
Vaidhyanathan, S. 2003. *Copyrights and copywrongs: The rise of intellectual property and how it threatens creativity.* New York: New York University Press.

Intellectual Property, History of

Zorina Khan

Abstract

The evolution of patents and copyrights followed different paths over time and across countries. Initially, intellectual property rules were endogenously determined according to social and economic priorities in each society. International patent laws subsequently were heavily influenced by early American policies that favoured the rights of original inventors. By contrast, US copyrights were among the weakest in the world; international copyright laws converged towards European doctrines that were based on non-economic rationales for inherent authors' rights. The intellectual property system in the 21st century therefore constitutes an anomaly, since previously no country simultaneously adhered to strong patent rights and strong copyrights.

Keywords

Barriers to entry; Copyright; Fair use; Innovation; Intellectual property; Intellectual property rights; Patents; Technology; Total factor productivity

JEL Classifications
N4

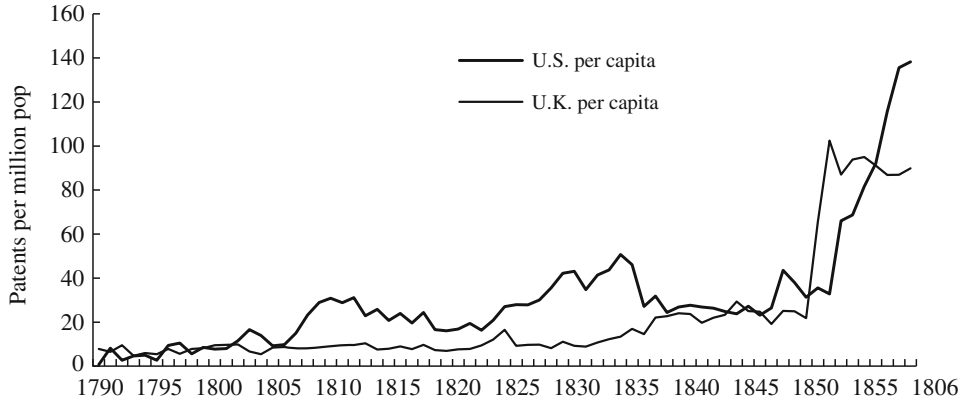
Intellectual property rights primarily have their origins in 15th-century monopoly privileges granted in Europe. Specific features of these rights of exclusion varied enormously and some constituted broad national claims that existed in perpetuity. By the 18th century, such differentiated privileges had evolved into standardized legal rights whose boundaries were delimited by statute. Most notably, the British Statute of Monopolies (1624) and the Statute of Anne (1710) established the longest continuous intellectual property system in existence. In Europe the philosophy and enforcement of intellectual property laws, the structure of patent and copyright systems, and the resulting patterns of invention (broadly defined to include technological and cultural creations) were all consistent with the oligarchic structure of these societies.

European patents were viewed as 'pernicious monopolies', which had to be narrowly interpreted, monitored, and restricted. This perspective was reinforced by the grant of patents to anyone who paid the exceedingly high fees, regardless of whether they were true inventors. The Crown reserved the right to expropriate any innovations that it wished, and kept others secret. Few provisions were made to ensure ready access to information. The legal system was biased against patents in general, and incremental improvements in particular. High transactions and monetary costs, as well as the prevailing prejudices towards non-elites, combined to create barriers to entry that discouraged the poor or disadvantaged from making contributions to technological innovation. Markets in patent rights and in patented inventions were thin and risky. As a result, trade secrecy probably played a more prominent part in protecting new discoveries, diffusion was certainly inhibited, the distribution of inventors and inventions was skewed, and potential inventors faced a great deal of uncertainty.

The elites who were privileged by these biases had little inducement to adopt institutional reforms that might generate social benefits at their expense. Administrators and patent agents lobbied against amendments and many had to be compensated for their lost rents before the system could be revised. Thus, despite their inefficiencies, patent rules and standards in both France and England remained essentially unchanged for stretches of over 100 years. In Britain, patent grants favoured a narrow range of capital-intensive industries and unbalanced growth paths. Clearly, despite these drawbacks, European economies still experienced industrialization and expansion; nevertheless, total factor productivity gains were quite modest and Britain was unable to sustain its initial advantage. Indeed, the record for Britain and other countries suggests that patent systems and their specific rules and standards had a significant effect. As Fig. 1 shows, when Britain reformed its laws in line with the United States in 1852 and 1883, patenting rates immediately increased. Similarly, Swiss patent reforms in the 1880s and Taiwanese revisions in the 1980s changed the rate and direction of their inventive activity (Khan 2005; Lo 2005).

In the United States policymakers were well aware of the European experience. They carefully weighed the grant of intellectual property rights against alternative strategies such as state subsidies and prizes. Legislators did not shrink from novel approaches, which they estimated would increase social welfare, regardless of how great the popular outcry. In accordance with the US Constitution, the utilitarian objective of the intellectual property system was to promote the public welfare. Patent and copyright laws were clearly distinguished in separate statutes in 1790, and developed along diametrically different lines based on a rational assessment of their costs and benefits.

The leading industrial nations acknowledged that patent rights might increase the rate of invention, but it was less conventional to propose that the background or the identity of inventors was irrelevant to their productivity. The US patent system exemplified one of the country's most democratic institutions, offering secure property rights to true inventors, regardless of colour, marital status, gender, or economic standing. Patent data, when linked to biographical information,



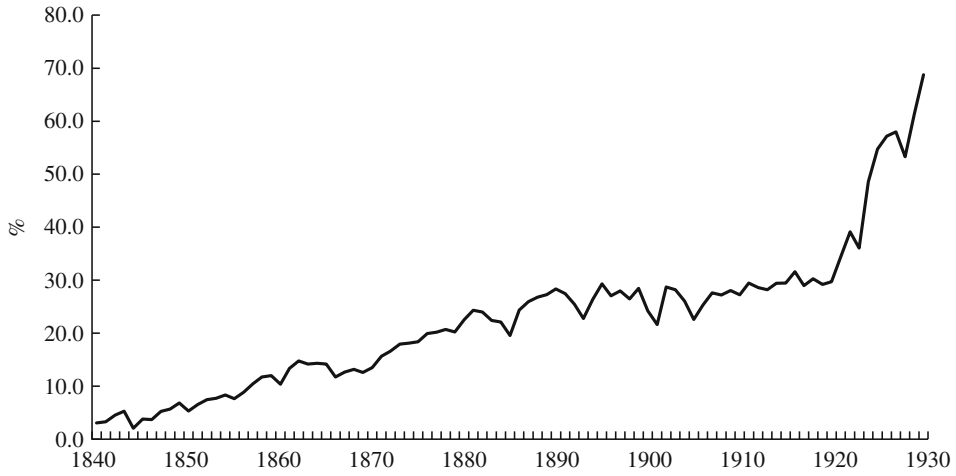
Intellectual Property, History of, Fig. 1 Patents per capita issued in Britain and the United States, 1790–1860 (*Source:* Khan (2005))

show that the expansion of markets and profit opportunities stimulated increases in inventive activity by attracting wider participation from relatively ordinary individuals. The roster of patentees included not only scientists and engineers, but also senators, schoolteachers, housewives, and even economists. The characteristics and patterns of patenting for American ' great inventors' were strikingly similar to those of ordinary patentees, unlike Europe where inventors were much more likely to be drawn from the elites.

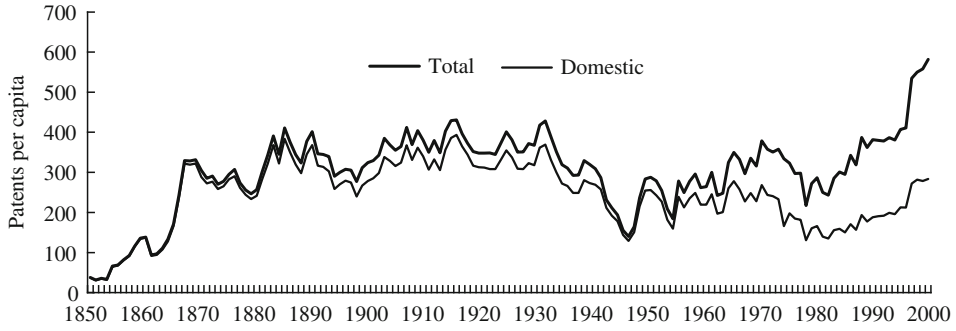
Such patterns were due in part to the conscious design of US patent institutions to ensure open access. These included transparent rules and administration, explicit measures for the diffusion of information, low fees, protection of the rights of the first and true inventor, a centralized examination system, and a legal system that balanced the rights of patentees with social welfare. American judges understood that secure private property rights and market competition comprised effective counters to oligarchical tendencies. Unlike the situation England, where the Crown reserved the right to expropriate inventions, in the United States even federal government claims could not trump the patentee's property right. The examination system ensured that all inventors were able to secure the services of professional examiners at minimal cost. Patents helped transform inventive ideas into tradeable assets, and this securitization of invention enhanced market efficiency.

The second industrial revolution from 1870 to 1920 was a transitional period that hinted at future changes in the nature and organization of technology. This era is usually characterized as the age of professional, science-based invention conducted by teams in research laboratories. Indeed, formal college education, human capital accumulation, and financial capital mobilization through corporate ties became more important, but relatively uneducated rural inventors were no less likely to produce valuable inventions. By the 1920s the rate of assignments sharply increased (Fig. 2), even as patenting per capita declined (Fig. 3), in part because inventive activity was increasingly conducted within corporations that appropriated returns through alternative strategies. Fig. 3 indicates that per capita patenting by US residents in the 21st century remains lower than during the second industrial revolution, whereas the so-called 'patent explosion' after the 1980s was largely the result of increases in grants to foreign inventors.

The US patent system was soon acknowledged as the most advanced in the world, and other countries drew causal connections between American achievements and its strong protection of patent property. Follower countries such as Germany and Japan patterned their own patent regime after the American model, but they introduced measures that addressed the particular needs of their own societies. These included the likelihood that patents would predominantly be granted to foreigners, the wish to raise revenues, and the



Intellectual Property, History of, Fig. 2 US patents assigned at time of issue as a proportion of all patents granted, 1840–1930 (*Source:* Khan (2005))



Intellectual Property, History of, Fig. 3 Per capita patenting: total US patents and patents issued to domestic residents, 1850–2000 (*Source:* US Patent Office)

need to foster domestic ingenuity. Their patent policies incorporated exemptions to protect social welfare in crucial industries such as food and pharmaceuticals, and restricted monopolistic tendencies through compulsory licensing and working requirements. Still, despite resistance from follower nations, patent harmonization over the 19th and 20th centuries converged towards the American model.

Copyrights

However much they praised and emulated US patent policies, other countries failed to understand the rationale for its copyright policies. The intellectual property clause of the US Constitution was the common source of both patent and copyright doctrines, and the same individuals were responsible for their formulation and implementation. American patent and copyright policies differed precisely because the objective of both systems was to promote the general welfare. This objective required a judicious balancing of private and public interests, the weighing of costs and benefits, and estimations of incentives and outcomes. Interests, costs and incentives differed across technical inventions and cultural goods, and also altered over time. Intellectual property adapted

endogenously to meet these changing circumstances in a way that contrasted directly with the institutional sclerosis in Europe.

The rationale for US copyrights was not based on European notions of inherent rights of personhood but, rather, on purely pragmatic and utilitarian grounds. Instead of a bona fide property right, American copyrights often mimicked more limited legal mechanisms such as contract, trade restraint or even liability rules. Americans viewed copyright trade-offs with greater concern. First, the economic processes that produced cultural goods differed from technological innovations: many copyrighted items might be produced even in the absence of financial incentives because their producers could benefit from ancillary returns such as enhanced reputations or greater demand for complementary goods. Second, the risk of unwarranted monopolies was higher, because cultural goods incorporated ideas that belonged to the public domain in ways that made it difficult to distinguish between the contributions of the author and those of society in general. Third, the enforcement of copyright had more serious implications for a democratic society. Restrictions on free diffusion could result in significant social costs in terms of knowledge, education and free speech, in ways that promised to bolster the narrow redistributive claims of elites and interest groups. Although policymakers protected property rights, their primary objective was not to benefit authors or publishing companies per se, so the advantages of a privileged few were circumscribed in order to protect the public domain.

It is, therefore, unsurprising that throughout US history patents were treated differently from copyrights. The first copyright statute granted protection to both ' authors and proprietors' for the instrumental purpose of learning, whereas only the first and true inventor could claim patent rights. Similarly, for much of the 19th century, work-for-hire doctrines led to weak employee rights in the case of copyrights, but not in the case of patents. Copyrights were administered in a registration system and were overturned if authors did not strictly comply with the rules; since 1836 patents were granted through an examination system and could not be revoked except for fraud. Patent policies were hostile to compulsory licences and unauthorized use of patent rights. By contrast, US copyright laws enshrine the world's most pervasive 'fair use' doctrines, which allow free unauthorized access for socially justifiable purposes (such as academic research and education) if such access did not significantly reduce the author's returns.

Although they excelled at pragmatic contrivances, 19th-century Americans were advisedly less sanguine about their efforts in the realm of music, art, literature, and drama, and so the USA was initially a net debtor in flows of material culture from Europe. The first copyright statute recognized this when it authorized international copyright piracy that persisted for a century. Proposals to reform the law were repeatedly brought before Congress and rejected because the net effects for Americans would be 'on the wrong side of the ledger'. It was only in 1891, when the balance of trade in cultural goods was more favourable to the United States, that an international copyright law was finally passed. Even then, the bill almost failed, and its passage required protectionist exemptions in favour of American workers and printing enterprises that remained in place until 1986. This policy was a dramatic departure from the evolution of international copyright laws in European countries. Early on in the 19th century France accorded national treatment to all countries, and led the movement for international harmonization towards strong copyright laws, which culminated in the 1886 Berne Convention. While it took a leadership role in patent conventions, the United Stated did not enter the Berne Convention until 1988, and it still has not completely complied with its provisions.

Today intellectual property rights are at the forefront of economic policy issues for developed and developing countries alike. Questions from four centuries ago are still current, ranging from the philosophical underpinnings of intellectual property to proposals for the abolition of all such rights. A 19th-century economist could assess contemporary policies that substituted tariffs and taxes for revenues to copyright owners, and would have

been equally familiar with analyses about whether uniformity in intellectual property rights across countries benefited global welfare. However, throughout their history, patent and copyright regimes have accommodated 'new eras' that were no less significant and contentious for their time than the 'digital dilemmas' of the 21st century.

Economic history indicates that intellectual property institutions best stimulated early economic growth when they enabled flexible endogenous responses to socioeconomic circumstances. However, the movement to harmonize patent and copyright laws encouraged a 'race to the top': it arose from two separate sources that culminated in stipulations for a system of uniformly strong patents and strong copyrights regardless of the level of economic development. Such a system did not exist anywhere in the world before the late-20th century, when countries enjoyed greater freedom to choose appropriate institutions. The more limited menu of choices today – especially for developing countries but even in the United States – constitutes an economic and historical anomaly.

See Also

- ▶ Intellectual Property
- ▶ Patents

Bibliography

Bugbee, B. 1967. *The genesis of American patent and copyright law*. Washington, DC: Public Affairs Press.

Dutton, H. 1984. *The patent system and inventive activity during the Industrial Revolution, 1750–1852*. Manchester: Manchester University Press.

Khan, B.Z. 2005. *The democratization of invention: Patents and copyrights in American economic development*. New York/Cambridge: NBER/Cambridge University Press.

Khan, B.Z., and K.L. Sokoloff. 2004. Institutions and democratic invention in 19th-century America. *American Economic Review* 94: 395–401.

Khan, B.Z., and K.L. Sokoloff. 2001. The early development of intellectual property institutions in the United States. *Journal of Economic Perspectives* 15(3): 233–246.

Lamoreaux, N.R., and K.L. Sokoloff. 2001. Market trade in patents and the rise of a class of specialized inventors in the nineteenth-century United States. *American Economic Review* 91: 39–44.

Lo, Shih-tse. 2005. Strengthening intellectual property rights: The experience of the 1986 Taiwanese patent reforms. Ph.D. thesis, University of California.

MacLeod, C. 1988. *Inventing the industrial revolution, the English Patent System, 1660–1800*. Cambridge: Cambridge University Press.

Sokoloff, K.L. 1988. Inventive activity in early industrial America: Evidence from patent records, 1790–1846. *Journal of Economic History* 48: 813–850.

Intelligence

Herbert Gintis

From the latter half of the 19th century to the Great Depression and the rise of fascism in the 1930s, it was fashionable both in and out of scientific circles to stress the contribution of the genetic worth of individuals and groups to their economic success. This stress was to be as often found among progressives, who used the doctrine to affirm birth control, divorce, and equal educational and economic opportunity for women, as among conservatives, who relied upon eugenic arguments to justify the natural superiority of their favoured social classes, ethic groups, and races. Eugenics, for instance, was supported by such radicals as Havelock Ellis, Beatrice and Sydney Webb and George Bernard Shaw, as well as such conservatives as Francis Galton, Leonard Darwin and Charles Davenport.

Brought into disrepute by its association with Nazism, the notion of genetic destiny resurfaced in the United States in the 1960s as a conservative reaction to the civil rights movement of American blacks (Jensen 1969; Herrnstein 1971; Eysenck 1971). The ensuing flurry of invective and empirical research has generated its proverbial quota of heat, and some light. My assessment of the evidence is that it provides no support for the notion that racial differences in economic success can be

attributed to their genetic inferiority with respect to mental functioning, since no acceptable technique of correcting for environmental differences between distinct racial groups has been devised. This same evidence provides some positive evidence for the effect of genes on economic performance in general, but it is so difficult to separate genetic and environmental factors, even among such relatively restricted samples as monozygotic (identical) twins, that the extent of this effect is unknown. It certainly is not enough to justify the use of the notion of genetic differences in any serious way in the formulation of economic policy.

To illustrate this point, consider one of the most careful and powerful examinations of the role of genes in explaining earnings (Taubman 1976a). Taubman uses a sample of 2468 pairs of monozygotic and dizygotic (fraternal) white male twins, attempting to explain differences in earnings at age 50 using such family background variables as parents' earnings and occupational status. This sample should provide the best possible evidence for or against the role of common genes, since monozygotic twins share all their genes, while dizygotic twins are no more genetically similar than two brothers. Assuming no assortive mating, no sex-linked genes, and no dominant and recessive genes, Taubman finds that the combined family environment explains 54% of the variance in earnings, while other influences explain the remaining 46%. However, depending on the extent to which twins share the same family environment more than two genetically unrelated individuals, the family contribution is apportioned so that the ratio of environmental to genetic factors ranges between 8% and 75%. In a related article (Taubman 1976b) using this sample, Taubman shows that not correcting for family background leads to a severe upward bias in estimating the returns to years of education. But the extent to which this bias is due to genetic as opposed to social factors cannot be ascertained.

It has been suggested (Jensen 1969) that individuals who perform badly on standardized cognitive tests be shunted out of the public educational system on grounds of the efficient application of economic resources. Certainly cognitive performance has been a central determinant of educational attainment in most modern societies. Yet one can show using a representative sample white American males (Bowles and Gintis 1976, pp. 110–12) that the economic return to education does not fall appreciably when cognitive performance is controlled in a regression analysis of earnings and occupational status. Moreover, it can be shown that for the same sample, the observed relationship between social class background and earnings is only in small part due to the tendency of families to pass on IQ differences (Bowles and Gintis 1976, pp. 120–22) either genetically or environmentally.

It is thus safe to say that if all differences in economic achievement were eliminated except for differences in IQ, and if the differences in the latter were maintained at their present level, there would be virtually perfect intergenerational economic mobility. In this sense, arguments which justify economic inequality, either among individuals or between races, on the basis of presumed intellectual differences, must be incorrect.

See Also

- ▶ Equality
- ▶ Human Capital
- ▶ Poverty

Bibliography

Bowles, S., and H. Gintis. 1976. *Schooling in capitalist America: Educational reform and the contradictions of economic life*. New York: Basic Books.

Eysenck, H. 1971. *The IQ argument*. New York: Modern Library.

Herrnstein, R. 1971. IQ. *Atlantic Monthly*, September.

Jensen, A. 1969. How much can we boost IQ and scholastic achievement? *Harvard Educational Review* 39(1): 1–123.

Taubman, P. 1976a. The determinants of earnings: genetics, family, and other environments; a study of white male twins. *American Economic Review* 66(5): 858–870.

Taubman, P. 1976b. Earnings, education, genetics, and environment. *Journal of Human Resources* 11(4): 447–461.

Interacting Agents in Finance

Cars Hommes

Abstract

Interacting agents in finance represent a behavioural, agent-based approach in which financial markets are viewed as complex adaptive systems consisting of many boundedly rational agents interacting through simple heterogeneous investment strategies, constantly adapting their behaviour in response to new information and strategy performance, and through social interactions. An interacting agent system acts as a noise filter, transforming and amplifying purely random news about economic fundamentals into an aggregate market outcome exhibiting important stylized facts such as unpredictable asset prices and returns, excess volatility, temporary bubbles and sudden crashes, large and persistent trading volume, clustered volatility and long memory.

Keywords

Agent-based modelling; Arbitrage; Asset prices; Asymmetric information; Behavioural finance; Bounded rationality; Bubbles; Clustered volatility; Discrete choice; Complexity; Evolutionary selection mechanisms; Excess volatility; Expectations; Finance; Herding; Cost of information; Interacting agents in finance; Long memory; Multiple equilibria; Noise traders; No-trade theorems; Random choice; Rational expectations; Representative agent; Social interaction; Stylized facts

JEL Classifications

G1; E3; D01; D84; D85

Interacting agents in finance represent a new behavioural, agent-based approach in which financial markets are viewed as complex adaptive systems consisting of many boundedly rational, heterogeneous agents interacting through simple investment strategies, constantly learning from each other as new information becomes available and adapting their behaviour accordingly over time. Simple interactions at the individual, micro level cause sophisticated structure and emergent phenomena at the aggregate, macro level. Recent surveys of this approach are Hommes (2006) and LeBaron (2006).

The traditional approach in finance is based on a representative, rational agent who makes optimal investment decisions and has rational expectations about future developments. Friedman (1953) made an early, strong argument in favour of rationality, arguing that 'irrational' agents would lose money whereas rational agents would earn higher profits. This is essentially an evolutionary argument saying that irrational agents will be driven out of the market by rational agents. In a perfectly rational world, information is transmitted instantaneously, asset prices reflect economic fundamentals and asset allocations are efficient. In the traditional view, agents interact only through the price system.

In contrast, Keynes earlier stressed that prices of speculative assets are not solely driven by market fundamentals, but that 'market psychology' also plays an important role. Another early critique on perfect rationality is due to Simon (1957), who emphasized that agents are limited in their computing abilities and face information gathering costs. Therefore individual behaviour is more accurately described by simple, suboptimal 'rules of thumb'. Along similar lines, Tversky and Kahneman (1974) in psychology argued that individual decision behaviour under uncertainty can be better described by simple *heuristics* and *biases.* Since the 1990s the traditional view of financial markets has been challenged through developments in bounded rationality (for example, Sargent 1993), behavioural finance (for example, Barberis and Thaler 2003) and computational, agent-based modelling (for example, Tesfatsion and Judd 2006).

Fundamentalists Versus Chartists

Most interacting agents models in finance include two important classes of investors:

fundamentalists and *chartists.* Fundamentalists base their investment decisions upon market fundamentals, such as interest rates, growth of the economy, company's earnings, and so on. Fundamentalists expect the asset price to move towards its fundamental value and buy (sell) assets that are undervalued (overvalued). In contrast, chartists or technical analysts look for simple patterns, for example, trends in past prices, and base their investment decisions upon extrapolation of these patterns. For a long time, technical analysis has been viewed as 'irrational' and, according to the Friedman argument, chartists would be driven out of the market by rational investors. Frankel and Froot (1986) were among the first to emphasize the role of fundamentalists and chartists in real financial markets. Evidence from survey data on exchange rate expectations (for example, Frankel and Froot 1987; Allen and Taylor 1990) shows that at short time horizons (say, up to three months) financial forecasters tend to use *destabilizing*, trend-following forecasting rules, whereas at longer horizons (say 3–12 months or longer) they tend to use *stabilizing,* mean- reverting, fundamental forecasts. Frankel and Froot (1986) argue that the interaction of chartists and fundamentalists amplified the strong rise and subsequent fall of the dollar exchange rate in the mid-1980s.

Another simple interacting agent system with chartists and fundamentalists driven by *herding* behaviour is due to Kirman (1991, 1993). This model was motivated by the puzzling behaviour of ants observed by entomologists. A colony of ants facing two identical food sources distributes asymmetrically, say 80–20 per cent, over the two sources. Moreover, at some point in time the distribution suddenly reverses to 20–80 per cent. Kirman (1993) proposed a simple stochastic model explaining ants' behaviour and applied it to a financial market setting (Kirman 1991). Agents can choose between two investment strategies – a fundamentalist or a chartist strategy– to invest in a risky asset. Two agents meet at random and with some interaction-conversion probability one agent will adopt the view of the other. There is also a small self-conversion probability that the agent will change her view no matter what the other agent believes. It turns out that, when the interaction- conversion probability is relatively high compared with the self-conversion probability, the distribution of agents is bimodal. The behaviour of the agents is very persistent and the market tends to be dominated by one group for a long time, but then the majority of agents suddenly switches to the other view, and so on.

But what about the Friedman argument? Will not 'irrational' technical trading rules be driven out of the market by rational investment strategies? DeLong et al. (1990) presented one of the first models showing that this need not be the case. Their model contains two types of traders, *noise traders,* with erroneous stochastic beliefs, and rational traders who are perfectly rational and take into account the presence of noise traders. Noise traders create extra risk and risk-averse rational traders are not willing to fully arbitrage away the mispricing. Noise traders bear more risk and can earn higher realized returns than rational traders, and therefore noise traders can survive in the long run. Lux (1995) presents a herding model with fundamentalists and chartists, whose behaviour is driven by imitation and past realized returns, leading to temporary bubbles and sudden crashes. Furthermore, Brock et al. (1992) showed empirically, using 90 years of daily Dow Jones index data, that technical trading rules can generate significant above-normal returns.

Markets as Complex Adaptive Systems

Since the end of the 1980s, multidisciplinary research as done at the Santa Fe Institute (SFI) (for example, Anderson et al. 1988) has stimulated a lot of work on interacting agents in economics and finance. Models of interacting particle systems in physics served as examples of how local interaction at the micro level may explain structure, for example a phase transition, at the macro level. This has motivated economists to study *the economy as an evolving complex system.*

Arthur et al. (1997) consider the so-called SFI artificial stock market consisting of an ocean of different types of agents choosing among many

simple investment strategies. Agents' investment decisions are affected by their expectations or beliefs about future asset prices. Beliefs affect realized prices, which in turn determine new beliefs, and so on. Prices and beliefs about prices thus *co-evolve* over time, and agents continuously adapt their behaviour as new observations become available, replacing less successful strategies by more successful ones. Are simple forecasting strategies irrational and will rational traders outperform technical traders in such an artificial market? In general, no. The reason is that a speculative asset market is an *expectations feedback system.* Imagine a situation where an asset price is overvalued and the majority of traders remains optimistic expecting the rising trend to continue. Aggregate demand will increase and as a result the asset price will rise even further. Optimistic expectations thus become *self-fulfilling* and chartists will earn higher realized returns than fundamental traders who sold or shortened the asset because they expected a decline in its price. As long as optimistic traders dominate the market and reinforce the price rise, fundamentalists will lose money. Even when the fundamentalists may be right in the long run, there are 'limits to arbitrage', for example due to short selling constraints, preventing them from holding their positions long enough against a prevailing optimistic view, as stressed by Shleifer and Vishny (1997).

Emergent Phenomena and Stylized Facts

The interacting agents approach has been strongly motivated by a number of important stylized facts observed in many financial time series (for example, Brock, 1997): (*a*) unpredictable asset prices and returns; (*b*) large, persistent trading volume; (*c*) excess volatility and persistent deviations from fundamental value, and (*d*) clustered volatility and long memory. According to (*a*) asset prices are difficult to predict. New information is absorbed quickly in asset prices and there is 'no easy free lunch', that is, arbitrage opportunities are difficult to find and exploit. The traditional rational, representative agent framework can explain (*a*), but has difficulty in explaining the other stylized facts (*b*)–(*d*). In particular, in a world with only rational, risk-averse investors with asymmetric information there can be no trade, because no trader can benefit from superior information since other rational traders will anticipate that this agent must have superior information and therefore will not agree to trade (for example, Fudenberg and Tirole 1991). These no-trade theorems are in sharp contrast to the huge daily trading volume observed in real financial markets, which suggests that there must be other types of heterogeneity such as *differences in opinion* about future movements. Stylized fact (*c*) means that fluctuations in asset prices are much larger than fluctuations in underlying market fundamentals. This point has been emphasized by, for example, by Shiller (1981). When markets are excessively volatile, prices can deviate from their fundamental values for a long time. Stylized fact (*d*) means that price fluctuations are characterized by irregular switching between quiet, low volatility phases, with small price fluctuations and turbulent phases of high volatility and large swings in asset prices. Interacting agent models have been able to explain these stylized facts simultaneously (for example, LeBaron et al. 1999; Lux and Marchesi 1999).

Evolutionary Selection of Strategies

Blume (1993) and Brock (1993) present a general probabilistic framework for strategy selection motivated by results from interacting particle systems in physics (see also Föllmer 1974). The probability of agents using strategy h changes over time according to a random utility fitness measure of the general form

$$U_{ht} = \pi_{ht} + S_{ht} + \varepsilon_{ht}. \tag{1}$$

Here π_{ht} represents *private utility,* for example given by (a weighted average of) realized profit, realized utility or forecasting performance. S_{ht} represents *social utility* measuring herding behaviour or social interactions (see Brock and Durlauf 2001a, b). For example, agents may behave as conformists, that is, they are more likely to follow

strategies that are more popular among the population (global interaction) or among their neighbours (local interaction). Agents observe the performance of each strategy with some idiosyncratic errors, represented by ε_{ht}.

A frequently used model for the probabilities or fractions of the different strategy types is the *discrete choice* or *multinomial logit model*

$$n_{ht} = e^{\beta U_{h,t-1}}/Z_{t-1}, \tag{2}$$

where $Z_{t-1} = \sum_j e^{\beta U_{j,t-1}}$ is a normalization factor so that the fractions add up to one. When the errors ε_{ht} in (1) are independently and identically distributed according to a double exponential distribution, the probability of choosing strategy h is exactly given by (2). The crucial feature of (2) is that, the higher the fitness of trading strategy h, the more agents will select strategy h, and therefore it is essentially an *evolutionary selection* mechanism. Agents are *boundedly rational* and tend to follow strategies that have performed well in the (recent) past. The parameter β is called the *intensity of choice* and is inversely related to the variance of the noise ε_{ht}. It measures how sensitive agents are to selecting the optimal strategy. The extreme case $\beta = 0$ corresponds to noise with infinite variance, so that differences in fitness cannot be observed and all fractions will be equal to 1/H, where H is the number of strategies. The other extreme $\beta = +\infty$ corresponds to the case without noise, so that the deterministic part of the fitness is observed perfectly, and in each period *all* agents choose the optimal forecast. An increase in the intensity of choice β represents an increase in the degree of rationality concerning strategy selection.

Brock and Hommes (1997, 1998) propose a simple, analytically tractable heterogeneous agent model to show how non-rational strategies can survive evolutionary selection. Brock and Hommes (1997) consider a market with an *endogenous evolutionary selection* of expectations rules described by the multinomial logit model (2), with fitness given by past realized profits. Agents choose between a set of different forecasting rules and tend to switch to forecasting strategies that have performed well in the recent past. When agents face information gathering costs, because sophisticated rational strategies are more costly to obtain, simple rule of thumb strategies can survive in this market. In Brock and Hommes (1998) this evolutionary selection of strategies is applied to a standard asset pricing model similar to but much simpler than the SFI artificial stock market. Agents choose between fundamentalists' and chartists' investment strategies. When the sensitivity to differences in past performance of the strategies is high (that is, the parameter β is high), evolutionary selection of strategies destabilizes the system and leads to complicated, possibly chaotic asset price fluctuations around the benchmark rational expectations fundamental price. The fluctuations are characterized by an irregular switching between a quiet phase with asset prices close to the fundamentals and a more turbulent phase with asset prices following (temporary) trends or bubbles. In contrast with Friedman's argument, chartists can survive in this evolutionary competition and may on average earn (short-run) profits equal to or even higher than (short-run) profits of fundamentalists.

A common finding in these models is that more rationality, that is, a larger intensity of choice, leads to *instability.* The intuition is that random choice leads to stability, because agents will be evenly distributed over the strategy space without systematic biases. In contrast, correlated choice may cause instability when, for example, many traders switch to a profitable trend-following strategy. Another common finding is that, when the social interaction effect is strong, multiple equilibria exist and it depends sensitively on the initial state to which of the many equilibria the market system will settle down (for example, Brock and Durlauf, 2001a, b).

Summary and Future Perspectives

Although the approach in finance is relatively new, interacting agent models have been able to explain important stylized facts simultaneously. An interacting agents system acts as a noise filter, transforming and amplifying purely random news about economic fundamentals into an aggregate

market outcome exhibiting excess volatility, temporary bubbles and sudden crashes, large and persistent trading volume, clustered volatility and long memory. It should be emphasized that at the aggregate level these asset price fluctuations are highly irregular and unpredictable, there exists no easy free lunch, and arbitrage will be very difficult and risky in such a market.

Much more theoretical work is needed in this area, for example, to find the 'simplest tractable model' explaining all important stylized facts. Speculative bubbles have been observed in laboratory experiments of Smith et al. (1988) and more recently in Hommes et al. (2005), showing that coordination on trend-following rules can destabilize a laboratory experimental asset market. Another important topic for future research is estimation of interacting agent models on financial data. Boswijk et al. (2007) is one of the first attempts to estimation of an evolutionary model with fundamentalists versus trend-following chartists using yearly S&P 500 data, suggesting that trend-following behaviour amplified the strong rise in stock prices at the end of the 1990s. More laboratory experiments and estimation of interacting agents models are needed to test the robustness and empirical relevance of the interacting agents approach.

See Also

- ▶ Economy as a Complex System
- ▶ Ergodicity and Nonergodicity in Economics
- ▶ Finance
- ▶ Finance (New Developments)
- ▶ Mathematics of Networks
- ▶ Network Formation
- ▶ Rationality, Bounded
- ▶ Residential Real Estate and Finance
- ▶ Social Interactions (Empirics)
- ▶ Social Interactions (Theory)

Bibliography

Allen, H., and M. Taylor. 1990. Charts, noise and fundamentals in the London foreign exchange market. *Economic Journal* 100: 49–59.

Anderson, P., K. Arrow, and D. Pines, eds. 1988. *The economy as an evolving complex system*. Reading: Addison-Wesley.

Arthur, W., J. Holland, B. LeBaron, R. Palmer, and P. Taylor. 1997. Asset pricing under endogenous expectations in an artificial stock market. In *The economy as an evolving complex system II*, ed. W. Arthur, S. Durlauf, and D. Lane. Reading: Addison-Wesley.

Barberis, N., and R. Thaler. 2003. A survey of behavioral finance. In *Handbook of the economics of finance*, ed. G. Constantinidis, M. Harris, and R. Stulz. Amsterdam: North-Holland.

Blume, L. 1993. The statistical mechanics of strategic interaction. *Games and economic behavior* 5: 387–424.

Boswijk, H., C. Hommes, and S. Manzan. 2007. Behavioral heterogeneity in stock prices. *Journal of Economic Dynamics and Control* (forthcoming).

Brock, W. 1993. Pathways to randomness in the economy: Emergent nonlinearity and chaos in economics and finance. *Estudios Económicos* 8: 3–55.

Brock, W. 1997. Asset price behavior in complex environments. In *The economy as an evolving complex system II*, ed. W. Arthur, S. Durlauf, and D. Lane. Reading: Addison-Wesley.

Brock, W., and S. Durlauf. 2001a. Discrete choice with social interactions. *Review of Economic Studies* 68: 235–260.

Brock, W., and S. Durlauf. 2001b. Interactions-based models. In *Handbook of econometrics*, ed. J. Heckman and E. Leamer, Vol. 5. Amsterdam: North-Holland.

Brock, W., and C. Hommes. 1997. A rational route to randomness. *Econometrica* 65: 1059–1095.

Brock, W., and C. Hommes. 1998. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control* 22: 1235–1274.

Brock, W., J. Lakonishok, and B. LeBaron. 1992. Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance* 47: 1731–1764.

DeLong, J., A. Shleifer, L. Summers, and R. Waldmann. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98: 703–738.

Föllmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1: 51–62.

Frankel, J., and K. Froot. 1986. Understanding the US dollar in the eighties: The expectations of chartists and fundamentalists. *Economic Record*, special issue, 24–38.

Frankel, J., and K. Froot. 1987. Using survey data to test standard propositions regarding exchange rate expectations. *American Economic Review* 77: 133–153.

Friedman, M. 1953. The case of flexible exchange rates. In *Essays in positive economics*. Chicago: University of Chicago Press.

Fudenberg, D., and J. Tirole. 1991. *Game theory*. Cambridge: MIT Press.

Hommes, C. 2006. Heterogeneous agent models in economics and finance. In *Handbook of computational economics, volume 2: Agent-based computational*

economics, ed. L. Tesfatsion and K. Judd, 1109–1186. Amsterdam: North-Holland.
Hommes, C., J. Sonnemans, J. Tuinstra, and H. van de Velden. 2005. Coordination of expectations in asset pricing experiments. *Review of Financial Studies* 18: 955–980.
Kirman, A. 1991. Epidemics of opinion and speculative bubbles in financial markets. In *Money and financial markets*, ed. M. Taylor. London: Macmillan.
Kirman, A. 1993. Ants, rationality and recruitment. *Quarterly Journal of Economics* 108: 137–156.
LeBaron, B. 2006. Agent-based computational finance. In *Handbook of computational economics, volume 2: Agent-based computational economics*, ed. L. Tesfatsion and K. Judd, 1187–1233. Amsterdam: North-Holland.
LeBaron, B., W. Arthur, and R. Palmer. 1999. Time series properties of an artificial stock market. *Journal of Economic Dynamics and Control* 23: 1487–1516.
Lux, T. 1995. Herd behavior, bubbles and crashes. *Economic Journal* 105: 881–896.
Lux, T., and M. Marchesi. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397: 498–500.
Sargent, T. 1993. *Bounded rationality in macroeconomics*. Oxford: Clarendon Press.
Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71: 421–436.
Shleifer, A., and R. Vishny. 1997. The limits of arbitrage. *Journal of Finance* 52: 35–55.
Simon, H. 1957. *Models of man*. New York: Wiley.
Smith, V., G. Suchanek, and A. Williams. 1988. Bubbles, crashes and endogenous expectations in experimental spot asset markets. *Econometrica* 56: 1119–1151.
Tesfatsion, L., and K. Judd, eds. 2006. *Handbook of computational economics, volume 2: Agent-based computational economics*. Amsterdam: North-Holland.
Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–1131.

Interdependent Preferences

Peter C. Fishburn

JEL Classifications
D7

Interdependent preferences arise in economic theory in the study of both individual decisions and group decisions. We imagine that a decision is required among alternatives in a set X and that the decision will depend on preferences between the elements in X. If the preferences represent different points of view about the relative desirability of the alternatives, of if they are based on multiple criteria that impinge on the decision, then we encounter the possibility of interdependent preferences.

There are two predominant approaches to interdependent preferences, the synthetic and the analytic. The synthetic approach begins with a set of preference relations on X and attempts to aggregate them into a holistic representative preference relation on X. This is done in social choice theory, where each original relation refers to the preferences of an individual in a social group. The aggregate relation is then referred to as a social preference relation. The synthetic approach also appears in studies of individual preferences, as when an individual rank-orders the alternatives for each of a number of criteria and then seeks a holistic ranking that combines the criteria rankings in a reasonable way.

In contrast, the analytic approach begins with a holistic preference relation on X and seeks to analyse its internal structure. This may involve a decomposition into components of preference, or it may concern trade-offs between factors that describe interactive contributions to overall preferences.

The synthetic approach often considers a list $(\gg_1, \gg_2, \ldots, \gg_n)$ of preference relations on X, where $x \gg_i y$ could mean that person i prefers x to y, or that an individual prefers x to y on the basis of criterion i. The problem may then be to specify a holistic relation $\gg = f(\gg_1, \gg_2, \ldots, \gg_n)$ for each possible n-tuple of individual relations.

The analytic approach often begins with X as a subset of the product $X_1 \times X_2 \times \ldots \times X_n$ of n other sets. It considers a holistic *is preferred to* relation $\gg$ on X and asks how $\gg$ depends on the X_i considered separately or in combination. Under suitably strong *independence* assumptions it may be possible to *define* $\gg_i$ for each i in a natural way from $\gg$ on X, and perhaps to establish a functional dependence of $\gg$ on the $\gg_i$. However, interdependencies among the factors will often preclude such a simple resolution.

Historical Remarks

During the rise of marginal utility analysis in the latter part of the 19th century (Stigler 1950), the utility of each commodity bundle in a set $X = X_1 \times X_2 \times \ldots \times X_n$ was thought of as an intuitively measurable quantity. Founders such as Jevons, Menger and Walras regarded x as preferred to y precisely when $u(x)$, the utility of x, is greater than $u(y)$. Their analytic approach ignored interdependencies since they used the independent additive utility form $u(x) = u_1(x_1) + \ldots + u_n(x_n)$.

Later writers such as Edgeworth, Fisher, Pareto and Slutsky discarded the additive decomposition for the general interdependent form $u(x_1, x_2, \ldots, x_n)$. Their ordinalist view of utilities as a mere reflection of a preference ordering remains dominant, and they considered interactive effects among goods, such as complementarities and substitutabilities. A fine example of interdependent analysis appears in Fisher (1892).

Fisher was also one of the first people to mention explicitly the interpersonal effect on individual utility (Stigler 1950, p. 324). This occurs when one's utility and consequent demand depend on other people's consumption and could generally be expressed by $u_i(x_1, \ldots, x_n)$ as consumer i's utility when x_j denotes the commodity bundle of consumer j. Pigou (1903) considered the interpersonal effect in modest detail, and Duesenberry (1949) explored it in greater depth, but it has never been a prominent concern in economic theory.

Early examples of the synthetic approach in social choice theory come from Borda and Condorcet in the late 1700s. They asked: Given a list of voter preference rankings on a set X of $m \geq 3$ nominees, what is the best way of selecting a winner? Borda's answer was to assign $m, m - 1, \ldots, 1$ points to each first, second,..., last place nominee in the rankings and to elect the nominee with the largest point total.

Condorcet advocated the election of a nominee who is preferred by a simple majority of voters to each other nominee in pairwise comparisons. Black (1958) contains an excellent review of their work and the proposals of later writers. The debate over good election methods continues today (Brams and Fishburn 1983).

The turning point for social choice theory was Arrow's (1951) discovery that a few appealing conditions for aggregating individual preference orders on three or more candidates into social preference orders were jointly incompatible. The avalanche of research set off by Arrow's discovery is represented in part by Sen (1970, 1977), Fishburn (1973), Pattanaik (1971), and Kelly (1978).

In the area of risky decision theory, we envision a risky alternative as a probability distribution x on potential outcomes in a set C and observe that such decisions involve multiple factors since they entail both chances and outcomes. Bernoulli (1738) argued that a reasonable person will choose a risky alternative from a set X of distributions that maximizes his expected utility $\sum x(c)\, u(c)$. He proposed that u be assessed without reference to chance since he held an intuitive measurability view of utility. Consequently, his approach is wholly synthetic.

Little changed in the foundations of risky decisions during the next two centuries. Then, in a complete turnabout, von Neumann and Morgenstern (1944) introduced the analytic approach by beginning with a preference relation $\gg$ on X. Axioms for $\gg$ on X were shown to imply the existence of a real valued function u on C such that, for all x and y in X, $x \gg y$ precisely when x has greater expected utility than y, and u is to be assessed on the basis of comparisons between distributions. With a few exceptions, most notably Allais (1953), subsequent research has adopted the von Neumann-Morgenstern approach.

In the rest of this essay we comment further on multiattribute preferences under 'certainty', interdependent preferences in risky decisions, and social choice theory.

Multiattribute Preferences

We assume throughout this section that $\gg$ is a strict preference relation on $X = X_1 \times X_2 \times \ldots \times X_n$. A given X_i could represent amounts of commodity i, consumption bundles available to

person i, levels of income and/or consumption in period i, or values that elements in X might have for criterion i. Also let u on X and u_i denote real valued functions.

A non-empty proper subset N of $\{1, 2, \ldots, n\}$ is defined to be $\gg$-*independent* if, for all x_N and y_N in the product of the X_i over N and for all $z_{(N)}$ in the product of the X_i over i *not* in N,

$$(xN, z(N)) \gg (yN, z(N)) \gg (xN, w(N)) \times \gg (yN, w(N)).$$

Most research for $\gg$ on X involves $\gg$-independence for some N, but this need not exclude elementary notions of preference interdependencies. Two models that presume all N to be $\gg$-independent are the additive model (see Krantz et al. 1971)

$$x \gg y \Leftrightarrow u_1(x_1) + \ldots + u_n(x_n) > u_1(y_1) + \ldots + u_n(y_n),$$

and the lexicographic model (Fishburn 1974a) that places a value hierarchy on the factors.

Relationships between factors in the additive model and the more general model $x \gg y \Leftrightarrow u(x) > u(y)$ with u continuous, are often characterized by indifference maps or iso-utility contours. Interdependence arises in the lexicographic model from the fact that a small change in one factor overwhelms all changes in factors that are lower in the hierarchy.

Situations in which only some of the N are $\gg$-independent are reviewed by Keeney and Raiffa (1976, ch. 3) and Krantz et al. (1971, ch. 7). Among other things, these models allow complete reversals in preferences over one factor at different fixed levels of the other factors. This, of course, is a very strong form of interdependence under which all N may fail to be $\gg$-independent.

Other general models for interdependent preferences are discussed by Fishburn (1972) for finite sets, and by Dyer and Sarin (1979) when u is viewed in the intuitive measurability way.

Models that explicitly incorporate the interpersonal effect in economic analysis have been investigated by Pollak (1976) and Wind (1976), among others. Pollak explores the influence of several versions of interdependence among individuals on short-run and long-run consumption within a group. Using models of demand that are locally linear in others' past consumption, he concludes that the distribution of income need not be a determinant of long-run per capita consumption patterns. Wind's work is representative of empirical approaches to the influence of others on an individual's choice behaviour.

Risky Decisions

Interdependent preferences in risky decisions fall into two categories. The first concerns special forms for $u(c) = (c_1, c_2, \ldots, c_n)$ in the context of von Neumann- Morgenstern expected utility theory when the outcome set C is a subset of a product set $C_1 \times C_2 \times \ldots \times C_n$. The second focuses on changes in the basic model that occur when the independence axiom that gives rise to the expected utility form $\sum x(c)u(c)$ is relaxed or dropped.

Decompositions of $u(c_1, c_2, \ldots, c_n)$ in the expected utility model have been axiomatized by various people. Reviews and extensions of much of this work appear in Keeney and Raiffa (1976) and Farquhar (1978). The simplest independent decompositions are the additive form and a multiplicative form. The first of these requires x and y to be indifferent whenever the marginal distributions of x and y on X_i are the same for every i. The multiplicative form arises when, for each non-empty proper subset N of $\{1, \ldots, n\}$, the preference order over marginal distributions on the product of the C_i for i in N, conditioned on fixed values of the other factors, does not depend on those fixed values.

An example of a more involved interdependent decomposition is the two-factor model (Fishburn and Farquhar 1982) $u(c_1, c_2) = f_1(c_1)g_1(c_2) + \ldots + f_m(c_1)g_m(c_2) + h(c_1)$, which clearly allows a variety of interactive effects.

In the basic formulation for expected utility, assume that X is closed under convex combinations $\lambda x + (1 - \lambda)y$ with $0 < \lambda < 1$ and x and y in

X. The *independence axiom* for expected utility asserts that, for all x, y and z in X and all $0 < \lambda < 1$.

$$x \gg y \Rightarrow \lambda x + (1 - \lambda)z \gg \lambda y + (1 - \lambda)z.$$

Systematic violations of this axiom uncovered in experiments by Allais (1953), Kahneman and Tversky (1979), and MacCrimmon and Larsson (1979) among others, have led to new theories of risky decisions (Kahneman and Tversky 1979; Machina 1982; Chew 1983; Fishburn 1982) that do not assume independence. Machina (1982) proposes a model that approximates expected utility locally but not globally. Fishburn (1982) weakens the usual transitivity and independence assumptions to obtain a non-separable model $x \gg y \Leftrightarrow \varphi(x, y) > 0$ that allows preference cycles.

Related interdependent generalizations of Savage's subjective expected utility model for decisions under uncertainty are developed by Loomes and Sugden (1982) and Schmeidler (1984).

Social Choice

Many problems in social choice theory are related to Condorcet's phenomenon of cyclical majorities. This phenomenon occurs when voters have transitive preferences yet every nominee is defeated by another nominee under simple majority comparisons. The simplest example has three nominees and three voters with $x \gg_1 y \gg_1 z$, $z \gg_2 y$ and $y \gg_3 z \gg_3 x$; x beats y, y beats z, and z beats x. Borda's point- summation procedure can fail to satisfy Condorcet's majority-choice principle, and it is notoriously sensitive to strategic voting. Moreover, all summation procedures based on decreasing weights for positions in voters' rankings are sensitive to nominees who have absolutely no chance of winning, but whose presence can affect the outcome.

Various problems and paradoxes for multicandidate elections that arise from combinatorial aspects of synthetic methods are discussed by Fishburn (1974b), Niemi and Riker (1976), Saari (1982) and Fishburn and Brams (1983). Analyses of strategic voting, which suggest that no sensible election method is immune from manipulation by falsification of preferences, are reviewed in Kelly (1978) and Pattanaik (1978).

Arrow's (1951) theorem offers a striking generalization of Condorcet's cyclical majorities phenomenon. Suppose X contains three or more nominees, each of n voters can have any preference ranking on X, and an aggregate ranking $\gg = f(\gg_1, \gg_2, \ldots, \gg_n)$ is desired for each list $(\gg_1, \gg_2, \ldots, \gg_n)$ of individual rankings. The question addressed by Arrow is whether there is any way of doing this that satisfies the following three conditions for all x and y in X:

1. Pareto optimality: if $x \gg_i y$ for all i, then $x > y$;
2. Binary independence: the aggregate preference between x and y depends solely on the voters' preferences between x and y;
3. Non-dictatorship: there is no i such that $x \gg y$ whenever $x \gg_i y$. Arrow's theorem says that it is impossible to satisfy all three conditions.

Several dozen related impossibility theorems have subsequently been developed by others. Many of these are noted in Kelly (1978) and Pattanaik (1978). As well as multi-profile theorems, like Arrow's, that use different lists of preference rankings to demonstrate impossibility, there are single-profile theorems (Roberts 1980) that use only one list with sufficient variety in the rankings to establish impossibility.

Impossibility theorems, voting paradoxes, and results on strategic manipulation highlight the difficulty of designing good election procedures. Recent research to alleviate such problems (Dasgupta et al. 1979; Laffont and Moulin 1982) focuses on the design of preference-revelation mechanisms (generalized ballots) and aggregation procedures that encourage people to vote in such a way that the outcome will agree with some theoretically best decision based on the true but unknown preferences of the voters. Other work, such as that on approval voting (Brams and Fishburn 1983), continues to search for simple synthetic methods that minimize the problems that beset these methods.

See Also

- ▶ Arrow's Theorem
- ▶ Externalities

Bibliography

Allais, M. 1953. Le comportement de l'homme rationnel devant de risque; critique des postulats et axiomes de l'école Américaine. *Econometrica* 21: 503–546.

Arrow, K.J. 1951. *Social choice and individual values*. New York: Wiley. 2nd ed., 1963.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientarium Imperialis Petropolitanae* 5: 175–192. Trans. by L. Sommer as 'Exposition of a new theory on the measurement of risk' *Econometrica* 22: 23–36, (1954).

Black, D. 1958. *The theory of committees and elections*. Cambridge: Cambridge University Press.

Brams, S.J., and P.C. Fishburn. 1983. *Approval voting*. Boston: Birkhäuser.

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.

Dasgupta, P., P. Hammond, and E. Maskin. 1979. The implementation of social choice rules: Some general results on incentive compatibility. *Review of Economic Studies* 46: 185–216.

Duesenberry, J.S. 1949. *Income, saving, and the theory of consumer behavior*. Cambridge, MA: Harvard University Press.

Dyer, J.S., and R.K. Sarin. 1979. Measurable multiattribute value functions. *Operations Research* 27: 810–822.

Farquhar, P.H. 1978. Multiple criteria problem solving. In *Interdependent criteria in utility analysis*, ed. S. Zionts, 131–180. Berlin: Springer.

Fishburn, P.C. 1972. Interdependent preferences on finite sets. *Journal of Mathematical Psychology* 9: 225–236.

Fishburn, P.C. 1973. *The theory of social choice*. Princeton: Princeton University Press.

Fishburn, P.C. 1974a. Lexicographic orders, utilities, and decision rules: A survey. *Management Science* 20: 1442–1471.

Fishburn, P.C. 1974b. Paradoxes of voting. *American Political Science Review* 68: 537–546.

Fishburn, P.C. 1982. Nontransitive measurable utility. *Journal of Mathematical Psychology* 26: 31–67.

Fishburn, P.C., and S.J. Brams. 1983. Paradoxes of preferential voting. *Mathematics Magazine* 56: 207–214.

Fishburn, P.C., and P.H. Farquhar. 1982. Finite-degree utility independence. *Mathematics of Operations Research* 7: 348–353.

Fisher, I. 1892. Mathematical investigations in the theory of values and prices. *Transactions of the Connecticut Academy of Arts and Sciences* 9: 1–124. Reprinted, New York: Augustus M. Kelley, 1965.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.

Keeney, R.L., and H. Raiffa. 1976. *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: Wiley.

Kelly, J.S. 1978. *Arrow impossibility theorems*. New York: Academic.

Krantz, D.H., R.D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of measurement. Volume I: Additive and polynomial representations*. New York: Academic.

Laffont, J.-J. Moulin, H., 1982. Special issue on implementation. Journal of Mathematical Economics 10(1).

Loomes, G., and R. Sugden. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal* 92: 805–824.

MacCrimmon, K.R., and S. Larsson. 1979. Utility theory: Axioms versus 'paradoxes'. In *Expected utility hypotheses and the Allais Paradox*, ed. M. Allais and O. Hagen, 333–409. Dordrecht: Reidel.

Machina, M.J. 1982. 'Expected utility' analysis without the independence axiom. *Econometrica* 50: 277–323.

Niemi, R.G., and W.H. Riker. 1976. The choice of voting systems. *Scientific American* 234: 21–27.

Pattanaik, P.K. 1971. *Voting and collective choice*. Cambridge: Cambridge University Press.

Pattanaik, P.K. 1978. *Strategy and group choice*. Amsterdam: North-Holland.

Pigou, A.C. 1903. Some remarks on utility. *Economic Journal* 13: 58–68.

Pollak, R.A. 1976. Interdependent preferences. *American Economic Review* 66: 309–320.

Roberts, K.W.S. 1980. Social choice theory: The single-profile and multi-profile approaches. *Review of Economic Studies* 47: 441–450.

Saari, D.G. 1982. Inconsistencies of weighted summation voting systems. *Mathematics of Operations Research* 7: 479–490.

Schmeidler, D. 1984. *Subjective probability and expected utility without additivity*. Preprint #84, Institute for Mathematics and its Application, University of Minnesota.

Sen, A.K. 1970. *Collective choice and social welfare*. San Francisco: Holden-Day.

Sen, A.K. 1977. Social choice theory: A re-examination. *Econometrica* 45: 53–89.

Stigler, G.J. 1950. The development of utility theory, part I and II. *Journal of Political Economy* 58 (307–27): 373–396.

von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press. 2nd ed., 1947; 3rd ed., 1953.

Wind, Y. 1976. Preference of relevant others and individual choice models. *Journal of Consumer Research* 3: 50–57.

Interest and Profit

Carlo Panico

The analysis of the relationship between the rates of interest and profit deals with how to integrate the theory of money with the theory of value and distribution. Different views on this subject reflect alternative positions expressed in the debates over these theories. They describe how changes in the financial markets affect the production process and vice versa, and have different policy implications.

The dominant view on the relation between the rates of interest and profits since Adam Smith has been that while monetary factors affect the *daily* variations of the 'money' or 'market' interest rate, in equilibrium the interest rate can only be equal to its 'average' or 'natural' or 'real' value (as alternatively named in the literature), which is determined, independently of monetary factors, by the same forces that determine the general rate of profits on the capital invested in the production process. This view, though dominant, is not the only one put forward in the literature. Indeed some outstanding economists (e.g., J.M. Keynes) have proposed alternative views according to which monetary factors are relevant, both temporarily and permanently, in determining the equilibrium level of economic variables.

The distinction between the notions of 'money' or 'market' interest rate and 'average' or 'natural' or 'real' interest rate can be traced in the writings of most economists dealing with this subject. Ricardo, for instance, drew a clear-cut distinction between these rates and provided a coherent analysis of how to relate them. The rate of profits was determined on the basis of a 'surplus' theory, in the tradition of the English classical political economists. He took as given the social product, the available technology and the real wage rate, disallowing any direct influence of monetary factors in the determination of the rate of profits. The 'average' or 'natural' interest rate was in his writings a *portion* of the rate of profits and was determined by the latter. As to the 'market' interest rate, it could undergo daily variations around its 'natural' level, on account of the changing conditions of competition between lenders and borrowers in the money markets.

Ricardo's writings clarify the role played by the analysis of the money markets in the treatment of this subject. He provided several examples to show how changes in the market interest rate occur and how this rate tends to move towards its natural level. The analysis of money markets, therefore, has to support the view that the rate of profits ultimately determines the interest rate by describing how competitive market forces make the natural level of the interest rate assert itself in the money markets.

Ricardo presented a coherent analysis of how the theory of the interest rate and of money has to be integrated with the classical theory of value and distribution, when the real wage rate is taken as given. Yet he did not provide a detailed analysis of the working of the money markets. Soon after his death some attempts to develop the latter analysis were presented by Tooke and J.S. Mill. Both used this analysis to criticize Ricardo's view and to claim that the interest rate can be determined both temporarily and permanently by causes which are independent of what happens to the rate of profits.

Tooke's and Mill's positions were stimulated by the sharp variations in the interest rate which occurred during and after the Napoleonic wars. These long-lasting variations in the interest rate were, according to them, the result of the policy followed to finance the Government debt, rather than the result of a change in the conditions of production implying a higher level of the rate of profits. The analysis of the interest rate they put forward had a strong influence on the economic literature. It failed, however, to give a correct account of the competitive market forces relating the rates of interest and profits. This led to the unconvincing view that the average interest rate and the rate of profits can move independently of each other.

A similar point of view was expressed some years later by Marx, who studied monetary issues at length both at a theoretical and at an empirical

level, with particular reference to the experience of the British financial system. By looking at this system, Marx developed the view that the most powerful pressure-groups operating in the financial markets were able to affect the interest rate permanently (and consequently their share of surplus-value produced) through the introduction of financial innovations and through their influence on State interventions regulating the legal and institutional arrangements of the financial markets.

To support this idea Marx presented a detailed analysis of the working of these markets and of the determination of both the average and the market interest rates in terms of supply and demand for liquid means. He stressed the need to reject the notion of a 'natural' rate of interest determined on the basis of technological or material laws of production, and pointed out the analytical conditions allowing a determination of the interest rate independent of the rate of profits and based on historical, conventional elements.

Yet, like Tooke and Mill in the 1820s, Marx failed to correctly relate these two rates. He maintained Ricardo's determination of the rate of profits in terms of a given real wage rate within a surplus theory of value and distribution, and failed to work out the effects of the operation of competitive market forces coming into action when a divergence between the rates of interest and profits comes about.

However, from Marx's writings some insights can be derived for analysing these forces, even if he did not carry them out himself. He pointed out that the banking sector, like the other industrial sectors, has to earn at least the general rate of profits on the wages and capital anticipated to carry on its activity. Changes in the interest rates affect the revenues (interest received on bank loans and financial assets) and the costs (which include payments for wages, interest on deposits and the rate of profits on the capital advanced) of the banking firms. This produces adjustment processes tending to restore the conditions of equilibrium between revenues and costs, which set some constraints linking the movements of the rates of interest and profits.

The economic literature of the years during which Marx was developing Ricardo's surplus theory of value and distribution shows the progressive abandonment of this theory, which implied an inverse relationship between the rate of profits and the real wage rate. The general trend in this literature was to determine instead these two rates independently of each other, as the specific contributions of capital and labour to production. Leading historians of economic thought have argued that it is not possible to claim that a *new* theory of value and distribution was actually presented in those years. The analyses were not clearly spelt out, particularly those which dealt with the concept of capital.

This new trend was reflected in the analysis of the relationship between the rates of interest and profits. The tendency prevailed to identify interest and profit and use them as synonyms. Tooke, Fullarton and James Wilson claimed in those years that a *permanent* variation in the interest rate affects the costs of production and the prices in the same direction. For them a permanent change in the interest rate was *the same thing* as a change in the rate of profits. No one spoke any longer of independent movements of these two rates. Indeed, the whole analysis of the relation between the average interest rate and the rate of profits faded away. The only issue left for discussion was the temporary fluctuations of the market interest rate. Besides, the abandonment of the surplus theory of value and distribution, the confusions as to the definition of capital, and the tendency to identify demand for and supply of *money*- or *banking*-capital with demand for and supply of *real* capital – all these made impossible at that time the development of a monetary theory of the rate of profits, of a theory, that is, which recognizes the influence of monetary factors on this rate.

The concept of capital and its analytical role were more precisely spelt out by the economists who introduced the marginalist or neoclassical theory of value and distribution in the 1870s and later. In this new theory too, the natural interest rate and the general rate of profits were *the same thing*. The money rate of interest could vary independently only temporarily. In equilibrium it had

to be equal to its natural value determined as the rate of return to be made on the *real* capital employed in production.

Walras explicitly stated that money markets, so relevant in the real world, are a 'superfetation' in marginalist theory. Later on, Wicksell, having presented a rather developed analysis of the role played by monetary factors in disequilibrium, concluded that the money interest rate depends in the last analysis upon the supply of and the demand for real capital. In equilibrium no room can be allowed for the action of monetary forces.

An example of the marginalist view of the relationship between the rates of interest and profits can be found in *A Treatise of Money*, published by Keynes in 1930. This book is based upon the marginalist separation between the 'real' department and the 'monetary' department of economics. In the real department, in line with the marginalist theory of value and distribution, the equilibrium or natural level of the distributive variables, the relative prices and the level of output (which turns out to be full employment) are simultaneously determined. In the monetary department, as analysed in *A Treatise on Money*, equilibrium values are taken as given (or rather known from the real department). The fluctuations of the price level are then analysed by looking at the variations in the evaluations of the expected yields of investment goods and at the fluctuations of the money interest rate around its natural level.

In *A Treatise on Money* the instability of the demand for investment and the analysis of liquidity preference are both present. The latter analysis, presented in this book in the form of 'bear and bull positions', describes the working of the money markets and how changes in the interest rate come about. Their presence, however, does not imply the abandonment of the marginalist approach, which asserts itself in the determination of the natural interest rate.

An alternative view was presented by Keynes a few years later in the *General Theory*. In this book Keynes took a critical attitude towards the marginalist theory. From 1932 on, he denied the validity of the separation between real and monetary departments, proposing instead a 'monetary theory of production', where monetary factors were directly relevant to determine the equilibrium level of output, of the interest rate and of other distributive variables. According to this view, the traditional causal relation between the rates of interest and profits was reversed. The level of the latter rate depended upon the former.

The introduction of the concept of a 'monetary theory of production' coincided in Keynes's writings with the abandonment of the concept of 'natural interest rate'. A new 'monetary' theory of the interest rate was instead proposed to determine the 'average' or 'durable' (as Keynes named it: 1936, p. 203) level of this rate. Presenting this theory, Keynes stressed its historical, conventional character by claiming that *any* level of interest which is accepted with sufficient conviction as *likely* to be durable, *will* be durable (Keynes 1936, p. 203). He pointed out that the policy of the monetary authority is a major determinant of the 'common opinion' as to the future value of the interest rate. But he also added that other elements of an economic or institutional character can affect this 'common opinion', for instance by persuading the public that the monetary authority will not be able to maintain its present policy.

However, the *General Theory* did not introduce any alternative analysis of how changes in the interest rate come about. The analysis of liquidity preference was represented in this book as dealing with the daily variations in the market interest rate and describing how the level of the interest rate, which is expected to be durable, tends to assert itself.

To support his new view as to the causal link between the rates of interest and profits, Keynes also presented in the *General Theory* an analysis of how competitive market forces tend to affect productive processes when temporary or persistent changes occur in the financial markets. This analysis was best framed in chapter 17 of this book. According to Keynes, investors in the financial and industrial sectors pay great attention to the rates of return of the different assets. They allow for certain differentials between these rates, which take into account the liquidity premium offered by the different assets. The equilibrium structure of the rates of return (that is the equilibrium differentials) is so determined. When

the actual differentials do not correspond to the equilibrium ones, competitive market forces come into action, producing adjustment processes which affect the prices of the assets. Autonomous variations of the interest rates – as Keynes argued in the *General Theory* – if persistent can thus cause changes in the same direction in the rate of profits.

The analysis of the competitive forces tending to relate the rates of interest and profits, proposed by Keynes in the *General Theory*, can be considered as complementary to that hinted by Marx and described above. They point out two different market mechanisms which tend to relate the movements of these two rates. Combined together, these two analyses provide a base to argue for a monetary determination of the rate of profits. Those who accept a historical conventional determination of the interest rate, can claim the existence of a causal relationship moving from this rate to the rate of profits. Monetary factors can therefore be directly allowed in the determination of the rate of profits.

Sraffa's recent rehabilitation of the surplus theory of value and distribution moves along these lines. Taking advantage of the possibility offered by this theory to determine one distributive variable independently of the others, he suggests that it is preferable in order to analyse the present conditions of capitalist economies, to consider the rate of profits as an independent variable (determined by the level of the rates of interest on money), instead of following the classical political economists of the last century who took the real wage rate as independently determined (Sraffa 1960, p. 33).

A new way to relate the rates of interest and profits, and consequently the theory of money, and that of value and distribution is therefore proposed within the recent rehabilitation of the surplus approach. This proposal can refer to the writings of outstanding economists to find theoretical support and to spell out its analytical implications.

The reconstruction of this analysis appears particularly relevant in the face of the present state of the neoclassical approach. As outstanding neoclassical economists have themselves recognized, no satisfactory integration between monetary and real variables has yet been presented within modern versions of neoclassical theory, that is, those developed after the work of Hicks (1939) and the subsequent works of Arrow, Debreu and Malinvaud. On account of this unsatisfactory integration, neoclassical economists still refer, as they themselves say, to the works of Wicksell and Fisher, that is, to those earlier versions of the neoclassical theory, whose internal consistency has been denied by the debate on capital theory of the 1960s and the 1970s.

The relationship between the rates of interest and profits can thus be considered one of the most open and controversial subjects of political economy.

See Also

- ▶ Equal Rates of Profit
- ▶ Profit and Profit Theory

Bibliography

Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.

Keynes, J.M. 1930. *A treatise on money*, vol. 2 vols. London: Macmillan.

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

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

Interest Rates

J. E. Ingersoll

Interest is payment for use of funds over a period of time, and the amount of interest paid per unit of time as a fraction of the balance is called the interest rate. In some contexts, economists have found it conceptually useful to refer to a single number, the interest rate. In fact, at any point in time there are many prevailing interest rates. The

rate actually charged will depend on such factors as the maturity of the loan, the credit-worthiness of the borrower, the amount of collateral, tax treatment of interest payments for both parties, and special features such as call provisions or sinking fund requirements.

A complete treatment of interest rates would account for all of these factors, but in fact it is hard enough to handle any one of them adequately. This entry considers one factor, the term to maturity for default-free bonds. This analysis of the *term structure of interest rates* will be approached from the partial equilibrium perspective of finance: the determinants of interest rates and the impact of changing short and long interest rates on the macroeconomy are not discussed.

Merton (1970 and other unpublished work) was the first to formulate the term structure problem using the continuous- time no-arbitrage framework exposited here. Cox et al. (1985a, b), Dothan (1978), Richard (1978), and Vasicek (1977) solved term structure problems of this type for different stochastic processes.

Interest Rates in a Certain Environment

Historically, the theory of interest rates has been burdened with cumbersome notation designed to distinguish among spot rates, forward rates and rates of different terms. Notation and terminology will be kept to a minimum here. The instantaneous spot rate of interest at time t for a loan to be repaid an instant later will be denoted by $r(t)$. $R(t, T)$ will denote the continuously compounded interest rate for a zero coupon bond sold at t to be repaid at T, and $P(t, T)$ will be the price or present value per \$1 face value of such a bond. The relation between these three quantities is

$$r(t) \equiv \lim_{T \downarrow t} R(t,T) \tag{1a}$$

$$P(t,T) \equiv e^{-R(t,T)(T-t)}. \tag{1b}$$

An investor who has funds to invest until time T could buy a T-period zero coupon bond with a guaranteed annualized return of $R(t, T)$. Alternatively, the investor could roll over a series of shorter bonds or buy a longer bond with the intention of selling it at time T. In the absence of uncertainty, all of these plans would have to realize the same final return to avoid the possibility of arbitrage. In particular

$$\frac{1}{P(t,T)} = \exp\left[\int_t^T r(s)\mathrm{d}s\right] \tag{2}$$

or the continuously compounded long rate must be the average of the instantaneous rates,

$$R(t,T) = \frac{1}{T-t}\int_t^T r(s)\mathrm{d}s. \tag{3}$$

(Note that with discrete compounding one plus the long rate is equal to the geometric mean of one plus the single-period rates. See Dybvig et al. (1986) for a catalogue of related results in both continuous and discrete time.)

One-Factor Models of Interest Rates in an Uncertain Environment

When future interest rates are not known in advance, these relations need not be realized, even on average, but the equilibrium that obtains will still depend on the trade-offs between different bond portfolio strategies. The resulting equilibrium relation among the interest rates will depend primarily on the information structure perceived by investors and, in particular, on the temporal resolution of uncertainty.

In this section we will assume that the information structure is Markov in the currently prevailing short rate, which is assumed to capture all currently available information relevant for pricing default-free bonds. $P(r, t, T)$ denotes the price at t of a zero coupon bond maturing at T with a face value of \$1, given that the currently prevailing short rate is r. The evolution of the interest rate is assumed to follow a diffusion process.

$$\mathrm{d}r = f(r,t)\mathrm{d}t + g(r,t)\mathrm{d}\omega. \tag{4}$$

Here $f(\cdot)$ measures the expected change in the interest rate per unit time, $g(\cdot)$ measures the standard deviation of changes in the interest rate per unit time, and $d\omega$ is the increment to a Wiener process.

The price of a zero coupon bond evolves according to

$$dP(r,t,T)/P(r,t,T) = \alpha(r,t,T)dt + \delta(r,t,T)d\omega \quad (5)$$

where $\alpha(\cdot)$ is the bond's (endogenous) instantaneous expected rate of return and $\delta(\cdot)$ is its instantaneous standard deviation. By Itô's lemma

$$\alpha(r,t,T)P(r,t,T) = \frac{1}{2}g^2(r,t)P_{rr} + f(r,t)P_r + p_t \quad (6a)$$

$$\delta(r,t,T)P(r,t,T) = g(r,t)P_r \quad (6b)$$

where subscripts denote partial differentiation.

Equation (6a) is a partial differential equation relating the prices of a given bond at different points of time and with different prevailing short rates. Together with the known value of a bond at its maturity [$P(r, T, T) = 1$] and mild regularity conditions, (6a) is equivalent to the integral

$$P(r,t,T) = E\left[\exp\left\{-\int_t^T \alpha[r(s),s,T]ds\right\}\right] \quad (7)$$

(Friedman 1975, Theorem 5.2). This integral demonstrates that when the source of uncertainty is a stochastically varying discount rate rather than a random cash flow, it is generally improper to discount by using the expected discount rate. Rather, we should use the geometric mean across states of the discount rate.

To price bonds using (6a) or (7), we must know $\alpha(\cdot)$. Intuitively, $\alpha(\cdot)$ will be equal to the risk-free rate plus a risk premium. As there is but a single source of uncertainty, the returns on all bonds will be perfectly correlated; therefore, the risk premium on any bond will be proportional to its exposure to the risk, and knowing $\alpha(\cdot)$ for one bond (for all interest rate levels) is sufficient. We can specify $\alpha(\cdot)$ by fiat, or it can be derived from an equilibrium model.

One equilibrium condition that is often imposed is the 'local' expectations hypothesis, $\alpha(r, t, T) = r$ (Cox et al. 1981). Under the local expectations hypothesis the partial differential equation for bond pricing derived from (6a) and the integral in (7) are fully determined given the distribution of interest rate changes. The local expectations hypothesis is a strong assumption, but for assest pricing purposes it is the only case that needs consideration. It can be shown that the absence of arbitrage implies that we can artificially reassign probabilities so that the local expectations hypothesis holds without changing any asset prices. The artificial probabilities are called the risk neutral probabilities or the equivalent martingale measure. Here is an informal proof in our context.

Consider any two zero coupon bonds. From (6a) their realized returns are perfectly correlated. Therefore, to ensure the absence of arbitrage possibilities, their expected excess returns must be proportional to their standard deviations

$$\begin{aligned}\alpha(r,t,T) - r &= \pi(r,t)\delta(r,t,T) \\ &= \pi(r,t)(P_r/P)g(r,t).\end{aligned} \quad (8)$$

The risk premium term $\pi(\cdot)$ cannot depend on the bond in question, which is why it does not depend on T. We can now write the equivalent diffusion process under the martingale measure. Because we can infer the variance of any diffusion from its sample path and because the martingale measure has the same set of possible events as the original probability measure, the martingale standard deviations must be the same i.e., $g(\cdot)$. The drift under the martingale measure must equate expected returns across assets. Therefore, the drift term under the equivalent martingale measure must be $f^*(r, t) \equiv f(r, t) - \pi(r, t)\, g(r, t)$. Using the martingale measure (6a) therefore becomes

$$rP = \frac{1}{2}g^2(r,t)P_{rr} + f^*(r,t)P_r + P_t \quad (9)$$

and its solution analogous to (7) is

$$P(r,t,T) = E^*\left\{\exp\left[-\int_t^T r(s),\mathrm{d}s\right]\right\}. \quad (10)$$

where $E^*[\cdot]$ denotes expectation under the modified process with f^* as the drift term.

To illustrate how these tools are used, consider the simplest model in Cox et al. (1985b). The stochastic process for the interest rate is

$$\mathrm{d}r = \kappa(\mu - r)\mathrm{d}t + \sigma\sqrt{r}\ \mathrm{d}\omega. \quad (11)$$

For this process, the interest rate is attracted elastically toward its mean value μ and is influenced by a noise term whose variance is proportional to the prevailing level of the interest rate. As a consequence, the interest rate cannot become negative. Assuming logarithmic utility, the drift term for the equivalent process is $f^* = \kappa\mu - (\kappa + \lambda)r$.

With this specification of f^*, (9) or (10) is solved by

$$P(r,t,T) = A(T-t)e^{-B(T-t)r} \quad (12)$$

where

$$A(\tau) \equiv \left[\frac{2\eta e^{(\kappa+\lambda+\eta)\tau/2}}{b(\tau)}\right]^{2\kappa\mu/\sigma^2}, B(\tau) \equiv \left[\frac{2(e^{\eta\tau}-1)}{b(\tau)}\right]$$
$$\times b(\tau) \equiv 2\eta + (\kappa+\lambda+\eta)(e^{\eta\tau}-1),$$
$$\eta \equiv \sqrt{(\kappa+\lambda)^2 + 2\sigma^2}.$$

The resulting zero coupon yield has a limit of

$$R_\infty = 2\kappa\mu/(\eta + \kappa + \lambda)$$

regardless of the current short rate. (The constancy of the long rate may seem like a severe restriction of this model. Actually, it is a property that is to be expected of any recurrent model. See Dybvig et al. 1986.) If the current short rate is below R_∞, then the yield curve is upward sloping. If the current rate is greater than μ then the yield curve slopes downward. For interest rates between these two levels the yield curve has a single hump.

Brown and Dybvig (1986) have estimated a reduced form of this model, yield curve by yield curve, to obtain a time series of parameter estimates. They found that the implied variance tracks actual interest rate volatility well. They also find some evidence of misspecification, including what appears to be a tax effect.

Bond Pricing with Multiple Sources of Uncertainty

One criticism of previous models is that they permit only a single source of uncertainty. As a result of this assumption all bonds are perfectly correlated and all yield curves are characterized by a single parameter. Typically there would be factors that influence long and short rates differently. However, the basic techniques for bond pricing remain the same.

Suppose for simplicity that interest rates are determined by the short rate and one other state variable, x. The assumed dynamics for the two state variables are

$$\begin{aligned} \mathrm{d}r &= f(r,x,t)\mathrm{d}t + g(r,x,t)\mathrm{d}\omega_1 \\ \mathrm{d}x &= \phi(r,x,t)\mathrm{d}t + \gamma(r,x,t)\mathrm{d}\omega_2. \end{aligned} \quad (13)$$

The resulting partial differential equation for bond pricing is

$$rP = \frac{1}{2}g^2P_{rr} + \rho g\gamma P_{rx} + \frac{1}{2}\gamma^2P_{xx} + f^*P_r + \phi^*P_x + P_t \quad (14)$$

where as before $f^*(\cdot)$ and $\varphi^*(\cdot)$ denote the modified (risk adjusted) drift terms under the equivalent martingale measure, and ρ is the correlation between the two Wiener processes. Solving this problem requires a specification of the joint stochastic process and the necessary modification of the equivalent martingale measure.

If the second state variable, x, is the price of some asset, then the risk premium, $-\pi_2(\cdot)\gamma(\cdot)$, associated with ω_2 is $rx - \varphi(\cdot)$ giving $\varphi^* = rx$ as in the Black–Scholes option pricing model. We can also infer the risk premium if x is functionally related to an asset's value and time. For example suppose that x is the yield-to-maturity on a zero coupon bond maturing at s. The price of this bond, $P(r, x, t, s)$, must itself satisfy the pricing equation.

As $P(r, x, t, s) \equiv \exp[-x(s-t)]$, we know all of its required partial derivatives. Substituting them into (14) and solving for $\varphi^*(\cdot)$ gives

$$\phi^*(r,x,t) = \frac{r - x - \frac{1}{2}\gamma^2(\cdot)(s-t)^2}{t-s} \tag{15}$$

(Ingersoll 1987). Of course, if x is not known to be related to a marketed asset, we can still specify the second risk premium arbitrarily or by reference to an equilibrium model.

Brennan and Schwartz (1979) used a finite difference numerical approximation to analyse a two factor model. In a test with US Government bonds, they concluded that the two factor model predicted bond prices much better than did a one factor model and on the whole was an adequate description of bond prices.

The traditional forecasting models using geometrically smoothed averages of past short rates as predictors of future rates can also be viewed as two (or more) factor models of this sort. (See Dobson et al. (1976) for a survey of the forecasting models.) For example, consider the model of Malkiel (1966) in which the short rate tends to return to a 'normal level', measured by a geometric average of past interest rates

$$x(t) \equiv \beta \int_0^\infty e^{-\beta s} r(t-s)\mathrm{d}s. \tag{16}$$

The dynamics of this model are

$$\mathrm{d}r = \mathsf{K}(x-r)\mathrm{d}t + \sigma(\cdot)\mathrm{d}\omega \tag{17a}$$

$$\mathrm{d}x = \beta(r-x)\mathrm{d}t. \tag{17b}$$

As changes in the state variable x are locally deterministic, no risk adjustment is required. The modification to the stochastic process for r is handled in the usual fashion.

No closed form solution is known for this model when $\sigma(\cdot) = \sigma\sqrt{r}$ as in the Cox et al. (1985b) model. A solution for this and similar problems with other lag structures is given in Cox et al. (1981) when the variance is a constant.

Another form of two factor model uses real interest rates and expected inflation as its two state variables (Richard 1978; Cox et al. 1985b). Besides providing more flexibility for a better empirical fit, this formulation permits an identification of real and nominal effects for separate consideration.

Applications of Interest Rate Models

This continuous-time no-arbitrage method of pricing bonds is an outgrowth of the option pricing literature which gives it certain advantages over more traditional approaches. One advantage is the provision of a fully specified model of bond prices for empirical work. Another major advantage of term structure models of this type is that they provide a framework for valuation that is consistent with all of our other models based on the absence of arbitrage.

Thus, in addition to pricing zero coupon bonds and determining the term structure, this method can handle any other interest rate valuation problem. For example Cox et al. (1985b) value call options on interest rates. Applications to futures contracts, variable rate instruments, mortgages, loan commitments, etc. have also been published. Equation (9) or a multiple factor version such as (14) remains the fundamental relation among interest rate contingent claims. To value a particular claim the appropriate boundary condition is used in place of $P(r, t, T) = 1$.

Another advantage of such models is that they give an explicit measure of the risk characteristics of the priced assets. These risk measures can be used in immunizing bond portfolios or in relative performance measurement. Because they are derived in models based on the absence of arbitrage they are not subject to the same criticisms that have been made of traditional duration measures (see Ingersoll et al. 1978).

Of theoretical interest is the relation between these models and the risk neutral or equivalent martingale valuation procedure. With a constant interest rate it can be shown that the value of a derivative asset that pays $H(S_T)$ at time T, contingent on the value S_T of its primitive asset and makes no other disbursements, is

$$V(S,t) = \mathrm{e}^{-r(T-t)E*}[H(S_T)]. \quad (18)$$

The expectation $E^*[\cdot]$ is taken with respect to the risk neutral process for the primitive's price under which the actual expected rate of return is replaced by the risk-free interest rate. With a stochastic interest rate, the valuation is

$$V(S,t) = E^*\left\{\exp\left[-\int_t^T r(s)\mathrm{d}s\right]H(S_T)\right\}. \quad (19)$$

This equation generalizes both equations (18) and (10). The expectation in (19) is over the joint distribution under the martingale measure of interest rate paths and S_T. That is, the expectation assumes that

$$\mathrm{d}S(t) = r(t)S\ \mathrm{d}t + \sigma(\cdot)S\ \mathrm{d}\omega_S \quad (20a)$$

$$\mathrm{d}r(t) = f(\cdot)\mathrm{d}t + g(\cdot)\mathrm{d}\omega_r \quad (20b)$$

plus the modified processes of any other state variables that determine the term structure.

See Also

- Arbitrage
- Continuous-Time Stochastic Processes
- Option Pricing Theory
- Term Structure of Interest Rates
- Wiener Process

Bibliography

Brennan, M.J., and E.S. Schwartz. 1979. A continuous time approach to the pricing of bonds. *Journal of Banking and Finance* 3(2): 133–155.

Brown, S.J., and P.H. Dybvig. 1986. The empirical investigation of the Cox, Ingersoll, Ross theory of the term structure of interest rates. *Journal of Finance* 41(3): 616–630.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1981. A re-examination of traditional hypotheses about the term structure of interest rates. *Journal of Finance* 36(4): 769–799.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1985a. An intertemporal general equilibrium model of asset prices. *Econometrica* 53(2): 363–384.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1985b. A theory of the term structure of interest rates. *Econometrica* 53(2): 385–407.

Dobson, S., R. Sutch, and D. Vanderford. 1976. An evaluation of alternative empirical models of the term structure of interest rates. *Journal of Finance* 31(4): 1035–1065.

Dothan, L.U. 1978. On the term structure of interest rates. *Journal of Financial Economics* 6(1): 59–69.

Dybvig, P.H., J.E. Ingersoll, and S.A. Ross. 1986. Long forward rates can never fall. Unpublished working paper, Yale University.

Friedman, A. 1975. *Stochastic differential equations and applications*, vol. 1. New York: Academic Press.

Ingersoll, J.E. 1987. *Theory of financial decision making*. Totowa, NJ: Rowman and Littlefield.

Ingersoll, J.E., J. Skelton, and R.L. Weil. 1978. Duration forty years later. *Journal of Financial and Quantitative Analysis* 13(4): 627–650.

Malkiel, B.G. 1966. *The term structure of interest rates: Expectations and behavior patterns*. Princeton: Princeton University Press.

Merton, R.C. 1970. A dynamic general equilibrium model of the asset market and its application to the pricing of the capital structure of the firm. Unpublished working paper, Sloan School of Management, MIT.

Richard, S.F. 1978. An arbitrage model of the term structure of interest rates. *Journal of Financial Economics* 6(1): 33–57.

Vasicek, O.A. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5(2): 177–188.

Interests

Albert O. Hirschman

'Interest' or 'interests' is one of the most central and controversial concepts in economics and, more generally, in social science and history. Since coming into widespread use in various European countries around the latter part of the 16th century as essentially the same Latinderived word (*intérêt, interesse*, etc.), the concept has stood for the fundamental forces, based on the drive for self-preservation and self-aggrandizement, that motivate or should motivate the actions of the prince or the state, of the individual, and later of groups of people occupying a similar social or economic position (classes, interest groups). When related to the individual, the concept has at times had a very inclusive

meaning, encompassing interest in honour, glory, self-respect, and even after-life, while at other times it became wholly confined to the drive for economic advantage. The esteem in which interest-motivated behaviour is held has also varied drastically. The term was originally pressed into service as a euphemism serving, already in the late Middle Ages, to make respectable an activity, the taking of interest on loans, that had long been considered contrary to divine law and known as the sin of usury. In its wider meanings, the term at times achieved enormous prestige as key to a workable and peaceful social order. But it has also been attacked as degrading to the human spirit and corrosive of the foundations of society. An inquiry into these multiple meanings and appreciations is in effect an exploration of much of economic history and in particular of the history of economic and political doctrine in the West over the past four centuries.

The concept, moreover, still plays a central role in contemporary economics and political economy: the construct of the self-interested, isolated individual who chooses freely and rationally between alternative courses of action after computing their prospective costs and benefits to him or herself, that is, while ignoring costs and benefits to other people and to society at large, underlies much of welfare economics; and the same perspective has yielded important, if disturbing contributions to a broader science of social interactions, such as the Prisoner's Dilemma theorem and the obstacles to collective action because of free riding.

Two essential elements appear to characterize interest-propelled action: *self-centredness*, that is, predominant attention of the actor to the consequences of any contemplated action for himself; and *rational calculation*, that is, a systematic attempt at evaluating prospective costs, benefits, satisfactions, and the like. Calculation could be considered the dominant element: once action is supposed to be informed only by careful estimation of costs and benefits, with most weight necessarily being given to those that are better known and more quantifiable, it tends to become self-referential by virtue of the simple fact that each person is best informed about his *own* satisfactions and disappointments.

Interest and Statecraft

Rational calculation also played the chief role in the emergence of the concept of interest-motivated action on the part of the prince in the 16th and 17th centuries. It accounts for the high marks interest-governed behaviour received during the late 16th- and early 17th-century phases of its career in politics. The term did duty on two fronts. First, it permitted the emergent science of statecraft to assimilate the important insights of Machiavelli. The author of *The Prince* had almost strained to advertise those aspects of politics that clashed with conventional morality. He dwelt on instances where the prince was well advised or even duty bound to practise cruelty, mendacity, treason, and so on. Just as, in connection with money lending, the term interest came into use as a euphemism for the earlier term usury, so did it impose itself on the political vocabulary as a means of anaesthetizing, assimilating and developing Machiavelli's shocking insights.

But in the early modern age, 'interest' was not only a label under which a ruler was given *new latitude* or was absolved from feeling guilty about following a practice that he had previously been taught to consider as immoral: the term also served to impose *new restraints* as it enjoined the Prince to pursue his interests with a rational, calculating spirit that would often imply prudence and moderation. At the beginning of the 17th century, the interests of the sovereign were contrasted with the wild and destructive passions, that is, with the immoderate and foolish seeking of glory and other excesses involved in pursuing the by then discredited heroic ideal of the Middle Ages and the Renaissance. This disciplinary aspect of the doctrine of interest was particularly driven home in the influential essay *On the Interest of Princes and States of Christendom* by the Huguenot statesman the Duke of Rohan (1579–1638).

The interest doctrine thus served to release the ruler from certain traditional restraints (or guilt

feelings) only to subject him to new ones that were felt to be far more efficacious than the well-worn appeals to religion, morals, or abstract reason. Genuine hope arose that, with princely or national interest as guide, statecraft would be able to produce a more stable political order and a more peaceful world.

Interest and Individual Behaviour

The early career of the interest concept with regard to statecraft finds a remarkable parallel in the role it played in shaping behaviour codes for individual men and women in society. Here also a new license went hand in hand with a new restraint.

The new license consisted in the legitimation and even praise that was bestowed upon the single-minded pursuit of material wealth and upon activities conducive to its accumulation. Just as Machiavelli had opened up new horizons for the Prince, so did Mandeville two centuries later list a number of 'don'ts' for the commoner, in this case primarily in relation to money making. Once again, a new insight into human behaviour or into the social order was first proclaimed as a startling, shocking paradox. Like Machiavelli, Mandeville presented his thesis on the beneficial effects on the general welfare of the luxury trades (which had long been strictly regulated) in the most scandalous possible fashion, by referring to the activities, drives and emotions associated with these trades as 'private vices'. Here again, his essential message was eventually absorbed into the general stock of accepted practice by changing the language with which he had proclaimed his discovery. For the third time, euphemistic resort was had to 'interest', this time in substitution for such terms as 'avarice', 'love of lucre', and so on. The transition from one set of terms to the other is reflected by the first lines of David Hume's essay 'On the Independency of Parliament':

> Political writers have established it as a maxim, that, in contriving any system of government and fixing the several checks and balances of the constitution, every man ought to be supposed a *knave*, and to have no other end, in all his actions, than private interest. By this interest we must govern him, and, by means of it, make him, notwithstanding his insatiable avarice and ambition, cooperate to public good (Hume, 1742, vol. I, pp. 117–18, emphasis in the original).

Here interest is explicitly equated with knavishness and 'insatiable avarice'. But soon thereafter the memory of these unsavoury synonyms of interest was suppressed, as in Adam Smith's famous statement about the butcher, the brewer, and the baker who are driven to supply us with our daily necessities through their interest rather than their benevolence. Smith thus did for Mandeville what the Duke of Rohan had done for Machiavelli. His doctrine of the Invisible Hand legitimated total absorption of the citizen in the pursuit of private gain and thereby served to assuage any guilt feelings that might have been harboured by the many Englishmen who were drawn into commerce and industry during the commercial expansion of the 18th century but had been brought up under the civic humanist code enjoining them to serve the public interest *directly* (Pocock 1982). They were now reassured that by pursuing gains they were doing so *indirectly*.

In fact, Adam Smith was not content to praise the pursuit of private gain. He also berated citizens' involvement in public affairs. Right after his Invisible Hand statement he wrote 'I have never known much good done by those who affected to trade for the public good' (1776, p. 423). Ten years before, Sir James Steuart had supplied an interesting explanation for a similar aversion toward citizens' involvement in public affairs.

> . . . were everyone to act for the public, and neglect himself, the statesman would be bewildered . . . were a people to become quite disinterested, there would be no possibility of governing them. Everyone might consider the interest of his country in a different light, and many might join in the ruin of it, by endeavouring to promote its advantages (1767, vol. I, pp. 243–44).

In counterpart to the new area of authorized and recommended behaviour, these statements point to the important *restraints* that accompanied the doctrine of interest. For the individual citizen or subject as for the ruler, interest-propelled action

meant originally action informed by rational calculation in any area of human activity – political, cultural, economic, personal and so on. In the 17th century and through part of the 18th, this sort of methodical, prudential, interest-guided action was seen as vastly preferable to actions dictated by the violent, unruly and disorderly passions. At the same time, the interests of the vast majority of people, that is of those outside of the highest reaches of power, came to be more narrowly defined as economic, material or 'moneyed' interests, probably because the non-elite was deemed to busy itself primarily with scrounging a living with no time left to worry about honour, glory, and the like. The infatuation with interest helped bestow legitimacy and prestige on commercial and related private activities, that had hitherto ranked rather low in public esteem; correspondingly, the Renaissance ideal of glory, with its implicit celebration of the public sphere, was downgraded and debunked as a mere exercise in the destructive passion of self-love (Hirschman 1977, pp. 31–42).

The Political Benefits of an Interest-Based Social Order

The idea that the interests, understood as the methodical pursuit and accumulation of private wealth, would bring a number of benefits in the political realm took various distinct forms. There was, first of all, the expectation that they would achieve at the macrolevel what they were supposed to accomplish for the individual: hold back the violent passions of the 'rulers of mankind'. Here the best-known proposition, voiced early in the 18th century, says that the expansion of commerce is incompatible with the use of force in international relations and would gradually make for a peaceful world. Still more utopian hopes were held out for the effects of commerce on domestic politics: the web of interests delicately woven by thousands of transactions would make it impossible for the sovereign to interpose his power brutally and wantonly through what was called '*grands coups d'autorité*' by Montesquieu or 'the folly of despotism' by Sir James Steuart. This thought was carried further in the early 19th century when the intricacies of expanding industrial production compounded those of commerce: in the technocratic vision of Saint-Simon the time was at hand when economic exigencies would put an end, not just to *abuses* of the power of the state, but to any power whatsoever of man over man: politics would be replaced by administration of 'things'. As is well known this conjecture was taken up by Marxism with its prediction of the withering away of the state under communism. An argument that a century earlier had been advanced on behalf of emergent capitalism was thus refurbished for a new, *anti*-capitalist utopia.

Another line of thought about the political effects of an interest-driven society looks less at the constraints such as society will impose upon those who govern than at the difficulties of the task of governing. As already noted, a world where people methodically pursue their private interests was believed to be far more predictable, and hence *more governable*, than one where the citizens are vying with each other for honour and glory.

The stability and lack of turbulence that were expected to characterize a country where men pursue singlemindedly their material interests were very much on the minds of some of the 'inventors' of America, such as James Madison and Alexander Hamilton. The enormous prestige and influence of the interest concept at the time of the founding of America is well expressed in Hamilton's statement:

> The safest reliance of every government is on man's interests. This is a principle of human nature, on which all political speculation, to be just, must be founded (Hamilton 1784), cited in Terence Ball 1983, p. 45).

Finally, a number of writers essentially extrapolated from the putative personality traits of the individual trader, as the prototype of interest-driven man, to the general characteristics of a society where traders would predominate. In the 18th century, perhaps as a result of some continuing disdain for economic pursuits, commerce and money-making were often described as essentially innocuous or 'innocent' pastimes, in

contrast no doubt with the more violent or more strenuous ways of the upper or lower classes. Commerce was to bring 'gentle' and 'polished' manners. In French, the term innocent appended to commerce was often coupled with *doux* (sweet, gentle) and what has been called the thesis of the *doux commerce* held that commerce was a powerful civilizing agent diffusing prudence, probity and similar virtues within and among trading societies (Hirschman 1977, 1982a). Only under the impact of the French Revolution did some doubt arise on the direction of the causal link between commerce and civilized society: taken aback by the outbreak of social violence on a large scale, Edmund Burke suggested that the expansion of commerce depended itself on the *prior* existence of 'manners' and 'civilization' and on what he called 'natural protecting principles' grounded in 'the spirit of a gentleman' and 'the spirit of religion' (Burke 1790, p. 115; Pocock 1982).

The Invisible Hand

The capstone of the doctrine of self-interest was of course Adam Smith's Invisible Hand. Even though this doctrine, being limited to the economic domain, was more modest than the earlier speculations on the beneficent *political* effects of trade and exchange, it soon came to dominate the discussion. An intriguing paradox was involved in stating that the *general* interest and welfare would be promoted by the self-interested activities of innumerable decentralized operators. To be sure, this was not the first nor the last time that such a claim of identity or coincidence or harmony of interests of a part with those of a whole has been put forward. Hobbes had advocated an absolute monarchy on the ground that this form of government brings about an identity of interest between ruler and ruled; as just noted, the writers of the Scottish Enlightenment saw an identity of interest between the general interests of British society and the interests of the middle ranks; such an identity between the interests of one class and those of society became later a cornerstone of Marxism, with the middling ranks having of course been supplanted by the proletariat; and finally, the American pluralist school in political science returned essentially to the Smithian scheme of harmony between many self-interests and the general interest, with Smith's individual economic operators having been replaced by contending 'interest groups' on the political stage.

All these *Harmonielehren* have two factors in common: the 'realistic' affirmation that we have to deal with men and women, or with groups thereof, 'as they really are', and an attempt to prove that it is possible to achieve a workable and progressive social order with these highly imperfect subjects, and, as it were, behind their backs. The mixture of paradoxical insight and alchemy involved in these constructs makes them powerfully attractive, but also accounts for their ultimate vulnerability.

The Interests Attacked

The 17th century was perhaps the real heyday of the interest doctrine. Governance of the social world by interest was then viewed as an alternative to the rule of destructive passions; that was surely a lesser evil, and possibly an outright blessing. In the 18th century, the doctrine received a substantial boost in the economic domain through the doctrine of the Invisible Hand, but it was indirectly weakened by the emergence of a more optimistic view of the passions: such passionate sentiments and emotions as curiosity, generosity, and sympathy were then given detailed attention, the latter in fact by Adam Smith himself in his *Theory of Moral Sentiments*. In comparison to such fine, newly discovered or rehabilitated springs of human action, interest no longer looked nearly so attractive. Here was one reason for the reaction against the interest paradigm that unfolded toward the end of the 18th century and was to fuel several powerful 19th-century intellectual movements.

Actually the passions did not have to be wholly transformed into benign sentiments to be thought respectable and even admirable by a new generation. Once the interests appeared to be truly in command with the vigorous commercial and industrial expansion of the age, a general lament

went up for 'the world we have lost'. The French Revolution brought another sense of loss and Edmund Burke joined the two when he exclaimed, in his *Reflections on the Revolution in France*, 'the age of chivalry is gone; that of sophisters, economists and calculators has succeeded; and the glory of Europe is extinguished for ever' (1790, p. 111). This famous statement came a bare 14 years after the *Wealth of Nations* had denounced the rule of the 'great lords' as a 'scene of violence, rapine and disorder' and had celebrated the benefits flowing from everyone catering to his interests through orderly economic pursuits. Now Burke was an intense admirer of Adam Smith and took much pride in the identity of views on economic matters between himself and Smith (Winch 1985; Himmelfarb 1984). His 'age of chivalry' statement, so contrary to the intellectual legacy of Smith, therefore signals one of those sudden changes in the general mood and understanding from one age to the next of which the exponents themselves are hardly aware. Burke's lament set the tone for much of the subsequent Romantic protest against an order based on the interests which, once it appeared to be dominant, was seen by many as lacking nobility, mystery, and beauty.

This nostalgic reaction merged with the observation that the interests, that is, the drive for material wealth, were not nearly as 'innocuous', 'innocent' or 'mild', as had been thought or advertised. To the contrary, it was now the drive for material advantage that suddenly loomed as a subversive force of enormous power. Thomas Carlyle thought that all traditional values were threatened by 'that brutish god-forgetting Profit-and-Loss Philosophy' and protested that 'cash payment is not the only nexus of man with man' (1843, p. 187). This phrase – cash-nexus – was taken over by Marx and Engels who used it to good effect in the first section of the *Communist Manifesto* where they painted a lurid picture of the moral and cultural havoc wrought by the conquering bourgeoisie.

Many other critics of capitalist society dwelt on the destructiveness of the new energies that were relased by a social order in which the interests were given free rein. In fact, the thought arose that these forces were so wild and out of control that they might undermine the very foundations on which the social order was resting, that they were thus bent on self-destruction. In a startling reversal, feudal society, which had earlier been treated as 'rude and barbarous' and was thought to be in permanent danger of dissolution because of the unchecked passions of violent rulers and grandees, was perceived in retrospect to have nurtured such values as honour, respect, friendship, trust and loyalty, that were essential for the functioning of an interest-dominated order, but were relentlessly, if inadvertently, undermined by it. This argument was already contained in part in Burke's assertion that it is civilized society that lays the groundwork for commerce rather than vice versa; it was elaborated by a large and diverse group of authors, from Richard Wagner via Schumpeter to Karl Polanyi and Fred Hirsch (Hirschman 1982a, pp. 1466–70).

The Interests Diluted

While the interest doctrine thus met with considerable opposition and criticism in the 19th century, its prestige remained nevertheless high, particularly because of the vigorous development of economics as a new body of scientific thought. Indeed, the success of this new science made for attempts to utilize its insights, such as the interest concept, for elucidating some non-economic aspects of the social world. In his *Essay on Government* (1820), James Mill formulated the first 'economic' theory of politics and based it – just as was later done by Schumpeter, Anthony Downs, Mancur Olson etc. – on the assumption of rational self-interest. But this widening of the use of the concept turned out to be something of a disservice. In politics, so Mill had to recognize, the gap between the 'real' interest of the citizen and 'a false supposition [i.e., perception] of interest' can be extremely wide and problematic (1820, p. 88). This difficulty provided an opening for Macaulay's withering attack in the *Edinburgh Review* (1829). Macaulay pointed out that Mill's theory was empty: interest 'means only that men,

if they can, will do as they choose . . . it is . . . idle to attribute any importance to a proposition which, when interpreted, means only that a man had rather do what he had rather do' (p. 125).

The charge that the interest doctrine was essentially tautological acquired greater force as more parties climbed on the bandwagon of interest, attempting to bend the concept to their own ends. As so many key concepts used in everyday discourse, 'interest' had never been strictly defined. While individual self-interest in material gain predominated, wider meanings were never completely lost sight of. An extremely wide and inclusive interpretation of the concept was put forward at a very early stage in its history: Pascal's Wager was nothing but an attempt to demonstrate that belief in God (hence, conduct in accordance with His precepts) was strictly in our (long-term) self-interest. Thus the concept of *enlightened* self-interest has a long history. But it received a boost and special, concrete meaning in the course of the 19th century. With the contemporary revolutionary outbreaks and movements as an ominous backdrop, advocates of social reform were able to argue that a dominant social group is well advised to surrender some of its privileges or to improve the plight of the lower classes so as to insure social peace. 'Enlightened' self-interest of the upper classes and conservative opinion was appealed to, for example, by the French and English advocates of universal suffrage or electoral reform at mid-century; it was similarly invoked by the promoters of the early social welfare legislation in Germany and elsewhere toward the end of the century, and again by Keynes and the Keynesians who favoured limited intervention of the state in the economy through countercyclical policy and 'automatic stabilizers' resulting from welfare state provisions. These appeals were often made by reformers who, while fully convinced of the intrinsic value and social justice of the measures they advocated, attempted to enlist the support of important groups by appealing to their 'longer-term' rather than short-term and therefore presumably *shortsighted* interests. But the advocacy was not only tactical. It was sincerely put forward and testified to the continued prestige of the notion that interest-motivated social behaviour was the best guarantee of a stable and harmonious social order.

Whereas enlightened self-interest was something the upper classes of society were in this manner pressed to ferret out, the lower classes were similarly exhorted, at about the same epoch but from different quarters, to raise their sights above day-to-day pursuits. Marx and the Marxists invited the working class to become aware of its *real interests* and to shed the 'false consciousness' from which it was said to be suffering as long as it did not throw itself wholeheartedly into the class struggle. Once again, the language of interests was borrowed for the purpose of characterizing and dignifying a type of behaviour a group was being pressed to adopt.

Here, then, was one way in which the concept of interest-motivated behaviour came to be diluted. Another was the progressive loss of the sharp distinction an earlier age had made between the passions and the interests. Already Adam Smith had used the two concepts jointly and interchangeably. Even though it became abundantly clear in the 19th century that the desire to accumulate wealth was anything but the 'calm passion' as which it had been commended by some 18th-century philosophers, there was no return to the earlier distinction between the interests and the passions or between the wild and the mild passions. Money-making had once and for all been identified with the concept of interest so that all forms of this activity, however passionate or irrational, were automatically thought of as interest-motivated. As striking new forms of accumulation and industrial or financial empire-building made their appearance, new concepts were introduced, such as entrepreneurial leadership and intuition (Schumpeter 1911) or the 'animal spirits' of the capitalists (Keynes 1936, pp. 161–63). But they were not contrasted with the interests, and were rather assumed to be one of their manifestations.

In this manner the interests came to cover virtually the entire range of human actions, from the narrowly self-centred to the sacrificially altruistic, and from the prudently calculated to the passionately compulsive. In the end, interest stood behind anything people do or wish to do and to explain human action by interest thus did turn into the

vacuous tautology denounced by Macaulay. At about the same time, other key and time-honoured concepts of economic analysis, such as utility and value, became similarly drained of their earlier psychological or normative content. The positivistically oriented science of economics that flourished during much of this century felt that it could do without any of these terms and replaced them by the less value- or psychology-laden 'revealed preference' and 'maximizing under constraints'. And thus it came to pass that interest, which had rendered such long and faithful service as a euphemism, was now superseded in turn by various even more neutral and colourless neologisms.

The development of the self-interest concept and of economic analysis in general in the direction of positivism and formalism may have been related to the discovery, toward the end of the 19th century, of the instinctual-intuitive, the habitual, the unconscious, the ideologically and neurotically driven–in short, to the extraordinary vogue for the nonrational that characterized virtually all of the influential philosophical, psychological and sociological thinking of the time. It was out of the question for economics, all based on rationally pursued self-interest, to incorporate the new findings into its own apparatus. So that discipline reacted to the contemporary intellectual temper by withdrawing from psychology to the greatest possible extent, by emptying its basic concepts of their psychological origin–a survival strategy that turned out to be highly successful. It is of course difficult to prove that the rise of the nonrational in psychology and sociology and the triumph of positivism and formalism in economics were truly connected in this way. Some evidence is supplied by the remarkable case of Pareto: he made fundamental contributions both to a sociology that stressed the complex 'non-logical' (as he put it) aspects of social action and to an economics that is emancipated from dependence on psychological hedonism.

Current Trends

Lately there have been signs of discontent with the progressive evisceration of the concept of interest. On the conservative side, there was a return to the orthodox meaning of interest and the doctrine of enlightened self-interest was impugned. Apart from the discovery, first made by Tocqueville, that reform is just as likely to unleash as to prevent revolution, it was pointed out that most well-meant reform moves and regulations have 'perverse' side effects which compound rather than alleviate the social ills one had set out to cure. It was best, so it appeared, not to stray from the narrow path of narrow self-interest, and it was confusing and pointless to dilute this concept.

Others agreed with the latter judgement, but for different reasons and with different conclusions. They also disliked the manoeuvre of having every kind of human action masquerade under the interest label. But they regarded as relevant for economics certain human actions and activities which cannot be accounted for by the traditional notion of self-interest: actions motivated by altruism, by commitment to ethical values, by concern for the group and the public interest, and, perhaps most important, the varieties of non-instrumental behaviour. A beginning has been made by various economists and other social scientists to take these kinds of activities seriously, that is, to abandon the attempt to categorize them as mere variants of interest-motivated activity (Boulding 1973; Collard 1978; Hirschman 1985; Margolis 1982; McPherson 1984; Phelps 1975; Schelling 1984; Sen 1977).

One important aspect of these various forms of behaviour which do not correspond to the classical concept of interest-motivated action is that they are subject to considerable variation. Take actions in the public interest as an example. There is a wide range of such actions, from total involvement in some protest movement down to voting on Election Day and further down to mere grumbling about, or commenting on, some public policy within a small circle of friends or family – what Guillermo O'Donnell has called 'horizontal voice' in contrast to the 'vertical' voice directly addressed to the authorities (1986). The actual degree of participation under more or less normal political conditions is subject to constant fluctuations along this continuum, in line with changes in economic conditions,

government performance, personal development, and many other factors. As a result, with total time for private *and* public activity being limited, the intensity of citizens' dedication to their private interests is also subject to constant change. Near-total privatization occurs only under certain authoritarian governments, for the most repressive regimes do not only do away with the free vote and any open manifestation of dissent, but also manage to suppress, through their display of terrorist power, all *private* expressions of inconformity with public policy, that is, all those manifestations of 'horizontal voice' that are actually important forms of public involvement.

An arresting conclusion follows. That vaunted ideal of predictability, that alleged idyll of a privatized citizenry paying busy and exclusive attention to its economic interests and thereby serving the public interest indirectly, but never directly, becomes a reality only under wholly nightmarish political conditions! More civilized political circumstances necessarily imply a less transparent and less predictable society.

Actually, this outcome of the current inquiries into activities not strictly motivated by traditional self-interest is all to the good: for the only certain and predictable feature of human affairs is their unpredictability and the futility of trying to reduce human action to a single motive–such as interest.

See Also

- ▶ Economic Interpretation of History
- ▶ Exit and Voice
- ▶ Invisible Hand
- ▶ Property
- ▶ Self-Interest

Bibliography

Ball, T. 1983. The ontological presuppositions and political consequences of a social science. In *Changing social science*, ed. D.R. Sabia Jr. and J.T. Wallulis. Albany: State University of New York Press.

Boulding, K.E. 1973. *The economy of love and fear: A preface to grants economics*. Belmont: Wadsworth.

Burke, E. 1790. *Reflections on the revolution in France*. Chicago: Regnery, 1955.

Carlyle, T. 1843. *Past and present*. New York: New York University Press, 1977.

Collard, D. 1978. *Altruism and economy: A study in non-selfish economics*. Oxford: Robertson.

Collini, S., D. Winch, and J. Burrow. 1983. *That noble science of politics: A study in nineteenth-century intellectual history*. Cambridge: Cambridge University Press.

Hamilton, A. 1784. Letters from Phocion, number I. In *The works of Alexander Hamilton*, ed. John C. Hamilton. New York: C.S. Francis, 1851, Vol. II, 322.

Himmelfarb, G. 1984. *The idea of poverty: England in the early industrial age*. New York: Knopf.

Hirschman, A.O. 1977. *The passions and the interests: Political arguments for capitalism before its triumph*. Princeton: Princeton University Press.

Hirschman, A.O. 1982a. Rival interpretations of market society: Civilizing, destructive, or feeble? *Journal of Economic Literature* 20(4): 1463–1484.

Hirschman, A.O. 1982b. *Shifting involvements: Private interest and public action*. Princeton: Princeton University Press.

Hirschman, A.O. 1985. Against parsimony: Three easy ways of complicating some categories of economic discourse. *Economics and Philosophy* 1: 7–21.

Hume, D. 1742. In *Essays moral, political and literary*, ed. T.H. Green and T.H. Grose. London: Longmans, 1898.

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

Macaulay, T.B. 1829. Mill's essay on government. In *Utilitarian logic and politics*, ed. J. Lively and J. Rees. Oxford: Clarendon, 1978.

McPherson, M.S. 1984. Limits on self-seeking: The role of morality in economic life. In *Neoclassical political economy*, ed. D.C. Colander, 71–85. Cambridge, MA: Ballinger.

Margolis, H. 1982. *Selfishness, altruism and rationality*. Cambridge: Cambridge University Press.

Meinecke, F. 1924. *Die Idee der Staatsräson in der Neueren Geschichte*. Munich: Oldenburg.

Mill, J. 1820. Essay on government. In *Utilitarian logic and politics*, ed. J. Lively and J. Rees. Oxford: Clarendon, 1978.

O'Donnell, G. 1986. On the convergences of Hirschman's exit, voice and loyalty and shifting involvements. In *Development, democracy and the art of trespassing: Essays in honor of A.O. Hirschman*, ed. A. Foxley et al. Notre Dame: University of Notre Dame Press.

Phelps, E.S. (ed.). 1975. *Altruism, morality and economic theory*. New York: Russell Sage Foundation.

Pocock, J.G.A. 1982. The political economy of Burke's analysis of the French revolution. *Historical Journal* 25: 331–349.

Rohan, H., Duc de. 1638. *De l'interêt des princes et états de la chrétiénité*. Paris: Pierre Margat.

Schelling, T.C. 1984. *Choice and consequence*. Cambridge, MA: Harvard University Press.

Schumpeter, J.A. 1911. *The theory of economic development*. Cambridge, MA: Harvard University Press, 1951.

I

Sen, A. 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6(4): 317–344.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. New York: Modern Library, 1937.
Steuart, J. 1767. In *Inquiry into the principles of political oeconomy*, ed. A.S. Skinner. Chicago: University of Chicago Press, 1966.
Winch, D. 1985. The Burke–Smith problem and late eighteenth century political and economic thought. *Historical Journal* 28(1): 231–247.

Intergenerational Income Mobility

Gary Solon

Abstract

One important aspect of income inequality is the extent to which position in the income distribution is passed from parents to children. Theoretical models suggest that both intergenerational persistence and equilibrium income inequality increase with the responsiveness of earnings to human capital investment and with the heritability of income-generating traits, and decrease with the progressivity of public investment in children's human capital. A rapidly growing empirical literature is documenting the extent of intergenerational income mobility in many countries and is beginning to explore why intergenerational transmission is as high (and low) as it is.

Keywords

Assortative mating; Becker, G.; Cobb–Douglas function; Human capital investment; Income mobility; Inequality of income; Intergenerational income mobility; Longitudinal surveys; Panel surveys

JEL Classifications

J62

'Intergenerational income mobility' refers to the degree to which position in the income distribution persists or changes from one generation to the next.

For example, a society in which individuals' adult income is altogether independent of their parents' income is a highly mobile society. A society in which one's percentile in the income distribution is always identical to one's parents' percentile is completely immobile. Neither extreme is ideal, and neither corresponds to the intergenerational patterns typically observed in actual societies. Which intergenerational association between the extremes is desirable depends on the processes generating it. In any case, reasonable and well-informed observers are likely to disagree about the optimal level of intergenerational mobility because of differences in their values, such as feelings about equity–efficiency trade-offs. See Jencks and Tach (2005) for a discussion of some of the normative issues.

Because understanding intergenerational mobility is important for understanding the nature of income inequality, intergenerational mobility has received a great deal of attention from both theoretical and empirical researchers. Since the late 1980s, empirical research describing the extent of intergenerational mobility has made considerable progress. Much work remains to improve our understanding of the causal processes underlying the observed extent of intergenerational mobility.

Theory

The classic theoretical analysis by Becker and Tomes (1979) encompasses a multitude of reasons why relative income status is correlated across generations. In a recent variant of that analysis, Solon (2004) adopts the functional form assumptions consistent with the log-linear intergenerational mobility regression commonly estimated by empirical researchers. In this model, an individual parent divides her income between her own consumption and investment in an individual child's human capital so as to maximize a Cobb–Douglas utility function in which the two

goods are the parent's consumption and the child's adult income. The mapping from the parent's investment in her child's human capital to the child's subsequent income as an adult operates through two functions.

First, a semi-logarithmic human capital production function relates the child's level of human capital to the logarithm of the sum of the parent's investment and public investment (for example, publicly supported education and health care for children) plus a variable representing the human capital endowments children receive regardless of the investment choices of their families and the government. These more mechanically determined endowments to children are, in Becker and Tomes's (1979, p. 1158) words, 'determined by the reputation and "connections" of their families, the contribution to the ability, race, and other characteristics of children from the genetic constitutions of their families, and the learning, skills, goals, and other "family commodities" acquired through belonging to a particular family culture'. The transmission of these endowments is assumed to follow a first-order autoregressive process across generations. Thus, intergenerational transmission occurs both because higher-income parents have greater wherewithal to invest in the human capital of their children and because of the genetic and cultural heritability of human capital. Second, the mapping to the child's income is completed by a semi- logarithmic earnings function that relates the child's log earnings to her level of human capital.

This simple model leaves out some important aspects of intergenerational transmission. For example, it assumes that the parent cannot borrow against the child's prospective earnings and does not bequeath financial assets to the child. See Becker and Tomes (1986) and Mulligan (1997) for analyses that relax this assumption. Also, the model's single-parent/single-child structure ignores the role of assortative mating, which is discussed by Lam and Schoeni (1994) and Chadwick and Solon (2002). Nevertheless, the model is rich enough to illustrate some key aspects of the intergenerational transmission process. These are embodied in the following result concerning the intergenerational income elasticity β, which is the coefficient in the regression of the child's log income on the parent's log income:

$$\beta \cong \frac{(1-\gamma)\theta + \lambda}{1 + (1-\gamma)\theta\lambda} \tag{1}$$

where θ is the elasticity of earnings with respect to human capital investment, λ is the autoregressive parameter representing the genetic and cultural heritability of income-generating traits, and γ is an index of the progressivity of public investment in children's human capital.

This result implies that the intergenerational income elasticity increases with the responsiveness of earnings to human capital investment and with the heritability of income-generating traits, and decreases with the progressivity of public investment in children's human capital. Cross-country differences in intergenerational mobility could arise from differences in any of these factors. So could changes over time in a particular country's intergenerational mobility. Finally, it can be shown that the same factors that increase the intergenerational income elasticity also increase the equilibrium level of cross-sectional income inequality within a generation. Thus, we should not be surprised if societies with particularly high income inequality also exhibit high intergenerational persistence of income status.

Empirical Evidence

If lifetime income data were available for both the parents' and children's generations in a nationally representative sample, estimation of the intergenerational income elasticity β could be performed simply by applying least squares to the regression of the children's log lifetime income on the parents' log lifetime income. In most countries, however, the ideal data are not available. As of the 1980s, data constraints had forced most of the then-small empirical literature to rely on short-term income measures, such as annual income in only one year, for peculiarly homogeneous samples. As summarized in Becker and Tomes (1986, p. S25), the resulting estimates

suggested that 'a 10% increase in father's earnings (or income) raises son's earnings by less than 2%'. As discussed in detail in Solon (1989, 1992), however, these estimates were biased substantially downward. The 'right-side' measurement error from using short-term parental income measures to proxy for parents' lifetime income can serve as a good classroom example for the econometrics textbook analysis of the attenuation bias resulting from 'noisy' measurement of an explanatory variable. And when the estimates were based on relatively homogeneous parent samples, this bias was aggravated by the diminished 'signal variance' in the explanatory variable.

By the 1990s, empirical researchers in the United States had the benefit of better data. By that time, two longitudinal surveys initiated in the late 1960s, the Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys (NLS) of labour market experience, had generated new data with an intergenerational span. Because these surveys used national probability samples, they were less subject to the problems from homogeneous samples. And because the longitudinal surveys repeatedly collected income information at each re-interview, they enabled exploration of the impact of using longer-term income measures. Many of the new studies, surveyed in Solon (1999), treated the errors-in-variables issue by averaging the parental income measure over several years. A typical finding was that, in a regression of son's log earnings on a multi-year measure of father's log earnings, the estimated slope coefficient was about 0.4 – that is, double the 0.2 value that previously had been described as an upper bound for the intergenerational elasticity. A few studies treated the errors-in-variables problem by performing instrumental variables estimation with parental characteristics like education or occupation used to instrument for measured parental income. That approach usually produced somewhat higher intergenerational elasticity estimates, but, as explained in Solon (1992), the consistency of such instrumental variables estimation depends on the 'excludability' of the instruments from the model for children's income.

Even the new estimates based on multi-year parental income data probably were too low. As emphasized in Solon (1992) and Mazumder (2005), averaging parental income over several years reduces but does not eliminate attenuation bias. Non-random attrition from the longitudinal surveys probably generated a weaker version of the sample homogeneity that had plagued earlier data-sets. And, as discussed in Reville (1995) and Haider and Solon (2006), many of the newer estimates have been biased by 'left-side' measurement error. At the time researchers began to use intergenerational data from the PSID and NLS, the offspring were only about 30 years old or even younger. For workers in their twenties, the log of current income as a proxy for log lifetime income is subject to 'mean-reverting' measurement error, instead of the classical measurement error typically analysed in econometrics textbooks. The mean reversion occurs because the workers who eventually will have high lifetime earnings typically experience steeper earnings growth. As a result, the early career gap in current earnings between workers with high and low lifetime earnings tends to understate their lifetime gap. This sort of mean-reverting measurement error in a dependent variable is still another source of attenuation bias. Once all these downward biases in the estimation of the intergenerational elasticity are considered, it becomes plausible that the intergenerational elasticity in the United States may well be as large as 0.5 or 0.6.

In recent years, researchers have estimated intergenerational elasticities for many other countries, sometimes with much larger samples than are available from the US surveys. As summarized in Solon (2002), the elasticity estimates for the United States and United Kingdom are towards the high end among developed countries, with considerably smaller estimates appearing for Canada, Sweden, Finland and Norway. Some new estimates for developing countries in Latin America (Dunn 2004; Ferreira and Veloso 2004; Grawe 2004) are even higher than the US and UK estimates. By and large, these cross-country comparisons accord with the theoretical prediction of greater intergenerational income persistence in countries with greater income inequality, higher returns to human capital, and less progressive public investment in children's human capital.

A related question is whether the changes in income inequality experienced by many countries since the 1970s have been accompanied by changing intergenerational elasticities. In most of the time-trends research conducted so far, the time spans and sample sizes have been too limited to permit strong conclusions.

Cross-country comparisons have only begun to illuminate *why* intergenerational income associations are as large (and as small) as they are. To what extent does intergenerational transmission occur because higher-income parents invest more in their children's human capital? What are the roles of genetic and cultural heritability? One intriguing line of research seeks clues from comparisons of relatives with varying degrees of genetic and environmental relatedness. Sibling studies of this type (Taubman 1976; Bjorklund et al. 2005) have compared correlations in socioeconomic status among monozygotic twins, dizygotic twins, non-twin full siblings, half-siblings, and biologically unrelated adoptive siblings, and also have compared biological siblings reared together and apart. (A related literature – Solon et al. 2000; Page and Solon 2003a, b; Oreopoulos 2003; Raaum et al. 2006 – has compared sibling correlations and correlations among unrelated children that grew up in the same neighbourhood. The typical finding that the sibling correlations are considerably larger than the neighbour correlations suggests that family influences loom larger than neighbourhood influences in accounting for the effects of origins on socioeconomic outcomes.) Intergenerational studies (Bjorklund et al. 2006; Plug 2004; Sacerdote 2004) have compared parent–child outcome associations in biological and adoptive families. Some empirical patterns consistent with an important role for genetic transmission are as follows: outcome correlations are particularly high among monozygotic twins; correlations for dizygotic twins and non-twin full siblings exceed those for half-siblings and adoptive siblings; correlations for biological siblings are positive even when the siblings are reared apart; intergenerational associations are higher for biologically related parents and children; and adoptive children's outcomes are positively associated with those of their biological parents (even after the adoptive parents' outcomes are controlled for). Empirical patterns consistent with important environmental factors are as follows: outcome correlations are positive among biologically unrelated adoptive siblings; correlations among biological siblings tend to be higher when the siblings are reared together; and adoptive children's outcomes are positively associated with those of their adoptive parents (even after the biological parents' outcomes are controlled for).

See Also

► Income Mobility

Bibliography

Becker, G., and N. Tomes. 1979. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87: 1153–1189.

Becker, G., and N. Tomes. 1986. Human capital and the rise and fall of families. *Journal of Labor Economics* 4: S1–S39.

Bjorklund, A., M. Jantti, and G. Solon. 2005. Influences of nature and nurture on earnings variation: A report on a study of various sibling types in Sweden. In *Unequal chances: Family background and economic success*, ed. S. Bowles, H. Gintis, and M. Osborne Groves. New York: Russell Sage Foundation.

Bjorklund, A., M. Lindahl, and E. Plug. 2006. The origins of intergenerational associations: Lessons from Swedish adoption data. *Quarterly Journal of Economics* 121: 999–1028.

Chadwick, L., and G. Solon. 2002. Intergenerational income mobility among daughters. *American Economic Review* 92: 335–344.

Dunn, C.E. 2004. *The intergenerational transmission of earnings: Evidence from Brazil*. Doctoral dissertation, University of Michigan.

Ferreira, S., and F. Veloso. 2004. *Intergenerational mobility of wages in Brazil*. Mimeo. Rio de Janeiro: IBMEC Business School.

Grawe, N. 2004. Intergenerational mobility for whom? The experience of high- and low-earning sons in international perspective. In *Generational income mobility in north America and Europe*, ed. M. Corak. Cambridge: Cambridge University Press.

Haider, S., and G. Solon. 2006. Life-cycle variation in the association between current and lifetime earnings. *American Economic Review* 96: 1308–1320.

Jencks, C., and L. Tach. 2005. Would equal opportunity mean more mobility? In *Mobility and inequality:*

I

Frontiers of research from sociology and economics, ed. S. Morgan, D. Grusky, and G. Fields. Stanford: Stanford University Press.
Lam, D., and R. Schoeni. 1994. Family ties and labor markets in the United States and Brazil. *Journal of Human Resources* 29: 1235–1258.
Mazumder, B. 2005. Fortunate sons: New estimates of intergenerational mobility in the U.S. using social security earnings data. *Review of Economics and Statistics* 87: 235–255.
Mulligan, C. 1997. *Parental priorities and economic inequality*. Chicago: University of Chicago Press.
Oreopoulos, P. 2003. The long-run consequences of living in a poor neighborhood. *Quarterly Journal of Economics* 118: 1533–1575.
Page, M., and G. Solon. 2003a. Correlations between brothers and neighboring boys in their adult earnings: The importance of being urban. *Journal of Labor Economics* 21: 831–855.
Page, M., and G. Solon. 2003b. Correlations between sisters and neighbouring girls in their subsequent income as adults. *Journal of Applied Econometrics* 18: 545–562.
Plug, E. 2004. Estimating the effect of mother's schooling on children's schooling using a sample of adoptees. *American Economic Review* 94: 358–368.
Raaum, O., K. Salvanes, and E. Sorensen. 2006. The neighbourhood is not what it used to be. *Economic Journal* 116: 200–222.
Reville, R. 1995. *Intertemporal and life cycle variation in measured intergenerational earnings mobility*. Mimeo. Santa Monica: RAND.
Sacerdote, B. 2004. *What happens when we randomly assign children to families?*, Working Paper No. 10894. Cambridge, MA: NBER.
Solon, G. 1989. Biases in the estimation of intergenerational earnings correlations. *Review of Economics and Statistics* 71: 172–174.
Solon, G. 1992. Intergenerational income mobility in the United States. *American Economic Review* 82: 393–408.
Solon, G. 1999. Intergenerational mobility in the labor market. In *Handbook of labor economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.
Solon, G. 2002. Cross-country differences in intergenerational earnings mobility. *Journal of Economic Perspectives* 16(3): 59–66.
Solon, G. 2004. A model of intergenerational mobility over time and place. In *Generational income mobility in north America and Europe*, ed. M. Corak. Cambridge: Cambridge University Press.
Solon, G., M. Page, and G. Duncan. 2000. Correlations between neighboring children in their subsequent educational attainment. *Review of Economics and Statistics* 82: 383–392.
Taubman, P. 1976. The determinants of earnings: Genetics, family, and other environments; a study of white male twins. *American Economic Review* 66: 858–870.

Intergenerational Models

Itzhak Zilcha

Recently it has been widely recognized by economists that the extension of the Arrow–Debreu model to a dynamic multi-period economy should not be restricted to models with a finite number of economic agents, each facing an infinite horizon. In analysing many economic problems, it seems natural to consider an open-ended horizon economy with individuals (or households) living for a finite number of periods; thus, at each date the economy consists of consumers of different ages (who interact with each other) and hence are inherently characterized by different economic parameters (such as their current income or planning horizon). In his seminal work, Samuelson (1958–9) attempts to explain, in an overlapping generations (OLG) equilibrium model, Irving Fisher's (1961) theory of interest. This simple model of a market economy characterized by an unbounded horizon, short-lived, overlapping, but essentially identical households, is different from the Arrow–Debreu model in various aspects. We shall concentrate upon similar models which have been successfully used to analyse microeconomic and macroeconomic problems. We shall focus upon the applications of these intergenerational models in: (1) efficient intergenerational and intertemporal allocation of resources, (2) intergenerational transfers (such as social security), and (3) optimal financing of government debt.

The Overlapping Generations Model: Pareto Optimality and Competitive Equilibria

Let us describe the OLG model as it has been frequently used in the literature. Basically this is a formal generalization of Samuelson's (1958) model.

The economy starts in period 1 and continues over periods extending indefinitely into the future, $t = 1, 2, \ldots$. In each period there are l, $l \geq 1$, consumption goods, which are perishable, and an imperishable fiat 'money'. All households (or consumers) in this economy live for two periods, except those households which already exist at the inception of the economy (namely, were born in period 0) and who live out the balance of their lives in period 1. The t th generation, G_t, is the set of all households born at the beginning of period t, $v = 0, 1, 2, \ldots$. Therefore, in each period t, there are just two age groups of households, the older generation, G_{t-1}, and the younger generation, G_t.

In each generation there are n households (or n 'types' of consumers). Each household will be denoted by a pair of indices (i, t), where $i = 1, \ldots, n$ and $t = 0, 1, \ldots$, and will be referred to as the ith household of the tth generation. Household (i, t), $t \geq 1$, has utility function u_i, u_i: $R^l_+ \times R^l_+ \to R$, defined over its lifetime consumption bundle $c_{it} = \left(c^y_{it}, c^0_{it}\right) \in R^l_+ \times R^l_+$, where c^y_{it} is the consumption of (i, t), when 'young', and c^0 is his consumption when 'old'. For $i \in G_0$ the utility function u_{i0} is defined over his consumption in period 1, c_{i0}, and u_{i0}: $R^l_+ \to R$. Each household (i, t) is endowed with physical goods in each period of his life $w_i = \left(w^y_i, w^0_i\right)$ in $R^l_+ \times R^l_+$ for $t \geq 1$. For $t = 0$, $w_{i0} \in R^l_+$ and $w_{i0} = w^0_i$. Households $(i, 0)$ of the oldest generation, G_0, may have initial endowments of money, $m_{i0} \geqq 0$.

Each consumer can trade goods and money in markets in t at (present value) prices denoted as $p_t \in R^l_+$ for goods and $p_m \in R^l_+$ for money, respectively. We also assume that young households can borrow money, free of transaction cost, by means of issuing IOUs as long as there are households in the same generation which will accept the IOUs. Each household (i, t) faces the problem, maximize $u_i(c_{it})$ s.t.

$$p_t c^y_{it} + p_{t+1} c^0_{it} \leqq p_t w^y_i + p_{t+1} w^0_i, \quad \text{for t} \geqq 1 \text{ and } c_{it} \geqq 0.$$

For i in G_0, the problem is maximize $u_{io}(c_{io})$ s.t.

$$p_1 c_{io} \leqq p_1 w_{io} + p_m m_{io}, \quad c_{io} \geq 0.$$

Note that since the economy is stationary, the optimal consumption of (i, t), $c_i(p_t, p_{t+1})$, depends only on (p_t, p_{t+1}), and for each (i, o) the optimal consumption $c_{io}(p_1, p_m)$ depends only on p_1 and p_m. The excess demand function for (i, t), $t \geq 1$, are

$$\begin{aligned} z_i\left(p_t, p_{t+1}\right) &= \left[z^y_i\left(p_t, p_{t+1}\right),\ z^0_i\left(p_t, p_{t+1}\right)\right] \\ &= c_i\left(p_t, p_{t+1}\right) - w_i. \end{aligned}$$

For (i, o) it is defined as $z_{io}(p_1, p_m) = c_{io}(p_1, p_m) - w_{io}$.

A competitive equilibrium is a triplet of positive goods prices, $(p^*_t)^\infty_{t=1}$ a non-negative price of money, p^*_m, and optimal lifetime consumption profiles $\left[(c^*_{it})^n_{i=1}\right]^\infty_{t=0}$ which are feasible, i.e.

(a)
$$\sum_{t=1}^{n} \left[c^{*y}_{it} + c^{*o}_{i(t-1)}\right] \leqq \sum_{t=1}^{n} \left[w^y_i + w^o_i\right], \quad \text{for } t = 1, 2, \ldots$$

and satisfies

(b)
$$c^*_{it} = z_i\left(p^*_i,\ p^*_{i+1}\right) + w_i, \quad \text{for all } i,\ t = 1, 2, \ldots,$$

(c)
$$c^*_{io} = z_{io}\left(p^*_1,\ p^*_m\right) + w_{io}, \quad \text{for all } i,$$

(d)
$$\sum_{t=1}^{n} \left[z^o_i\left(p^*_{t-1},\ p^*_t\right) + Z^y_i\left(p^*_t,\ p^*_{t+1}\right)\right] = 0 \qquad \text{for all } t = 2, 3, \ldots,$$

(e)
$$\sum_{t=1}^{n} \left[Z_{io}\left(p^*_1,\ p^*_m\right) + z^y_i\left(p^*_1,\ p^*_2\right)\right] = 0$$

A competitive equilibrium is called a *monetary equilibrium* if $p^*_m = 0$, and a *non-monetary* equilibrium, if $p^*_m = 0$.

A feasible allocation $\left((c_{it})^n_{i=1}\right)^\infty_{t=0}$ is called *Pareto optimal* (P.O.) if there is no other feasible

allocation $\left((\hat{c}_{it})_{i=1}^{n}\right)_{t=0}^{\infty}$ such that $u_i(\hat{c}_{it}) \geq u_i(c_{it})$ for all (i, t) and for some (h, τ), $1 \leqq h \leqq n$, $u_h(\hat{c}_{h\tau}) > u_h(c_{h\tau})$.

Samuelson analysed stationary allocations only and showed that without some extra-market institution, such as the fiat money already introduced in our presentation, competitive equilibria may not be Pareto optimal. Consider for example the case $l = 1$ $n = 1$, where the initial endowments are $w_i = (3, 1)$. Let $u_i(c^y{}_i, c^0{}_i) = \text{In } c_{io}$. Without fiat money the initial endowments allocation is the only competitive equilibrium when $p_t = 3^t$ for all t. This allocation is not P.O. since it is dominated by the P.O. allocation: $c^*_{i0} = 2$ and $c^*_i = (2, 2)$ for all i. If G_0 'contrives' fiat money, then a monetary equilibrium exists with this allocation and $p^*_t = 1$ for all t.

The main difference between an Arrow–Debreu model and the OLG model is the 'double infinity' of economic agents and commodities in the latter case. The reason for the failure of the second Theorem of Welfate Economics in the OLG models is the non–validity of Walras's Law in these infinite horizon models. This inherent property is due to the fact that, unlike the Arrow–Debreu economy, there is no posibility of trade between generations which do not overlap in their lifetime periods (see Gale 1973).

Extending Samuelson's analysis to the non–-stationary economies it was demonstrated by Cass et al. (1979) that when households are heterogeneous the following situations may occur: (a) there are both barter and monetary equilibria which are Pareto optimal; (b) there are both barter and monetary equilibria but none which is Pareto optimal. The non-optimality case, (b), highlights a fundamental difficulty: the mere creation of fiat money, i.e. the once-and-for-all augmentation of initial wealth, does not imply the second basic theorem of welfare economics; just the presence of money (trading for commodities at any conceivable positive price) may possibly not guarantee the Pareto optimality of competitive equilibrium.

In a non–stationary model, Okuno and Zilcha (1980) obtained a complete characterization of Pareto optimal competitive equilibria using the equilibrium prices. Also it is shown that certain monetary transfers (e.g. when the stock of money is increased at a uniform positive rate) can never achieve a Pareto efficient competitive equilibrium. Similar results have been attained by Balasko and Shell (1981). Cass and Shell (1983) introduced the concept of sunspot equilibria, that is equilibria in which uncertainty extrinsic to the economy operates through expectations to yield a fulfilled expectations competitive equilibrium in which the extrinsic randomness has real effects on prices and allocations. Cass and Shell examine assumptions on market structure and dynamics and find that sunspots can have no influence in a static world with complete markets, but can have effects when an OLG dynamics is introduced.

Intergenerational Transfers

The impact of annuity markets and social security upon savings, bequest and consumption has been studied recently in OLG models with uncertain lifetime. In these models there is a *continuum* of households in each generation. Each household maximizes its lifetime expected utility, where the utility from bequests appears explicitly. Uncertainty about lifetime concerns either the retirement period (as in Sheshinski and Weiss 1981; Eckstein et al. 1985) or may extend to all periods (as in Karni and Zilcha 1986). In the latter case it affects lifetime *earnings* and thus life insurance plays an important role in achieving efficient intergenerational allocations. In this model the first period income of each household depends upon the history of his family. In the presence of social security programmes and annuity markets agents can share death-related risks. When annuity markets operate, a non-discriminatory social security programme affects only the intergenerational allocation of resources. In the absence of private information regarding the survival probabilities, such a programme may lead to a non-optimal intragenerational allocation.

Kotlikoff and Summers (1981) use US data to estimate directly the contribution of intergenerational transfers to aggregate capital accumulation. They show that intergenerational

transfers account for the vast majority of aggregate US capital formation; only a negligible fraction of actual capital accumulation can be traced to life-cycle savings.

Loury (1981) models the dynamics of earnings distribution among successive generations of workers as a stochastic process. The process arises from random assignments of abilities to individuals by nature together with the utility maximizing bequest decisions of their parents. In this OLG model parents cannot borrow to make human capital investments in their offspring. Consequently the allocation of training resources among the young generation depends upon the distribution of earnings among their parents. This implies in turn that the conflict between egalitarian and redistributive policies and economic efficiency is mitigated.

National Debt: Analysis in Olg Models

In recent years the OLG model has been applied to investigate the effects of national debt on real economic activity. Historically, attention has been focused on the question of whether or not individuals perceive government bonds as net wealth, the link between wealth and real activity being as given. Diamond (1965) analyses these questions in an OLG model with production (employing a durable capital good), in which individuals provide for their retirement years by lending to entrepreneurs. Diamond's framework has been used extensively in analysing optimal financing of government expenditures and debt; we shall describe it briefly here. Individuals live for two periods: working period and retirement period. The labour force in each period t, $t = 0, 1, 2, \ldots$, is $L_t = L_0 (1 + n)^t$. The output in period t is given by $Y_t = F(K_t, L_t)$ where K_t is the capital stock in period t. $K_t + Y_t$, should be divided between capital stock K_{t+1} available for production in the date $t + 1$ and aggregate consumption C_t. Unlike the optimal growth models, where the central planner determines the allocation between productive capital and consumption in each date (according to some social welfare function) Diamond considers allocations through the competitive mechanism. Individuals in G_t receive wage W_t which equals the marginal product of labour, $F_L(K_t, L_t)$. This wage is allocated between current consumption $c^y{}_t$ and future consumption $C^o{}_{t+1}$ given the rate of interest on one period loans r_{t+1}, in a way that maximizes his lifetime utility $U\left(c_t^y, c_{t+1}^o\right)$. Therefore, $c_t^y = w_t - s_t$, $c_{t+1}^o = (1 + r_{r=1})s_t$ where $s_t = s(w_t, r_{r+1})$ is given by $U_1 = (1 + r_{t+1})U_2$. The equilibrium interest rate will equal the marginal product of capital, $r_{t+1} = F_K(K_{t+1}, L_{t+1})$, where the capital stock K_{t+1}, is *the sum* of the individual's savings $S_t = L_t s(w_t, r_{t+1})$. The equilibrium condition in the capital market which relates the interest rate to the wage rate of the previous period is ($f(c)$ is the per-capita production function);

$$r_{r+1} = f'\left(\frac{S_t}{L_{t+1}}\right) = f'\left[\frac{s(w_t, r_{t+1}}{1+n}\right] \qquad (1)$$

Given the factor–price frontier $w_t = \varphi(r_t)$ and the initial (w_1, r_1) equations (1) define the equilibrium path. Diamond observes from these relations that the equilibrium need not occur at an interest rate exceeding the Golden Rule level. Thus the competitive solution may be dynamically inefficient. The open-ended nature of this economy is crucial to such analysis. Using this OLG model, Diamond demonstrates that external debt has two effects in the long run, both arising from the taxes needed to finance the interest payments. The taxes directly reduce available lifetime consumption of the individual taxpayer. Further, by reducing his disposable income taxes reduce his savings and thus the capital stock. Internal debt has both of these effects as well as a further reduction in the capital stock arising from the substitution of government debt for physical capital in individual portfolios. Barro (1974), using the same model, assumes that generation t's utility depends on its own consumption and leisure and upon the utility of its immediate offspring's utility. This actually connects generation t to all future generations since its utility depends on the entire future time path of consumption and leisure of its descendants. Barro argued that despite the

limitation of finite lives of consumers in each generation, bonds will not be regarded as net wealth in a system characterized by intergenerational transfers. However, Barro admits that his neutrality result is valid for bonds that are redeemed at a known date. For a growing economy an increase in steady-state per-capita debt will generate net wealth if the rate of growth is greater than the interest rate. Further results concerning national debt were attained by Buiter (1979) and McCallum (1984).

Lucas (1972) uses Diamond's OLG framework, without population growth but including fiat money issued by the government and transferred to the old generation. Lucas shows that equilibrium prices and quantities exhibit what may be the central feature of the modern business cycle: a systematic relation between the rate of change in nominal prices and the level of real output. The relationship, essentially a variant of the well-known Phillips curve, is derived within a framework from which all forms of 'money illusion' are rigorously excluded: all prices are market clearing, all agents behave optimally in light of their objectives and expectations are formed optimally.

Tirole (1985) explores the interaction between productive and nonproductive savings in the model described above with capital accumulation. Some consequences of asset bubbles to asset pricing are derived. Once again the intergenerational interaction in an open-ended economy is important to such analysis.

Intergenerational models have been used in various other domains in economic theory due to their special dynamic characteristics. In international trade, for example, the examination of several questions related to exchange rate regimes has been carried out in an OLG model. Kareken and Wallace (1977) examine in various exchange rate regimes the differences that monetary-fiscal policy make. In the absence of capital controls the equilibrium exchange rate of the floating rate regime is indeterminate.

Bental (1985) applied an OLG model to two countries, two goods and two factors of production to show that in some cases the laissez-faire regime with free international trade and capital mobility is not necessarily a Pareto improvement over a regime in which capital is not internationally mobile. Furthermore, despite the fact that the laissez-faire regime is Pareto optimal and the restricted (portfolio autarky) regime is not, there do not exist simple temporal or intertemporal tax-transfer schemes which render the first allocation Pareto superior to the second.

See Also

- ▶ Overlapping Generations Model of General Equilibrium
- ▶ Social Security
- ▶ Sunspot Equilibrium

Bibliography

Balasko, Y., and K. Shell. 1981. The overlapping generations model II: The case of pure exchange with money. *Journal of Economic Theory* 24(1): 112–142.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(6): 1095–1117.

Bental, B. 1985. Is capital mobility always desirable? A welfare analysis of portfolio autarky in a growing economy. *International Economic Review* 26(1): 203–212.

Buiter, W. 1979. Government finance in an overlapping generations model with gifts and bequests. In *Social security versus private savings*, ed. G.M. von Furstenberg. Cambridge: Ballinger.

Cass, D., and K. Shell. 1983. Do sunspots matter? *Journal of Political Economy* 91(2): 193–227.

Cass, D., M. Okuno, and I. Zilcha. 1979. The role of money in supporting the Pareto optimality of competitive equilibria in consumption loan models. *Journal of Economic Theory* 20(1): 41–80.

Diamond, P.A. 1965. National debt in a neoclassical growth model. *American Economic Review* 55(5): 1126–1150.

Eckstein, Z., M.S. Eichenbaum, and D. Peled. 1985. Uncertain lifetimes and the welfare enhancing properties of annuity markets and social security. *Journal of Public Economics* 26(3): 303–326.

Gale, D. 1973. Pure exchange equilibrium of dynamic economic models. *Journal of Economic Theory* 6(1): 12–36.

Kareken, J.H., and N. Wallace. 1977. Portfolio autarky: A welfare analysis. *Journal of International Economics* 7(1): 19–43.

Karni, E., and I. Zilcha. 1986. Welfare and comparative statics implications of fair social security: A steady state analysis. *Journal of Public Economics* 30(1): 341–357.

Kotlikoff, L.J., and L.H. Summers. 1981. The role of intergenerational transfers in aggregate capital accumulation. *Journal of Political Economy* 89(4): 706–732.
Loury, G.C. 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49(4): 843–867.
Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4(2): 103–124.
McCallum, B. 1984. Are bond-financed deficits inflationary? A Ricardian analysis. *Journal of Political Economy* 92(1): 123–135.
Okuno, M., and I. Zilcha. 1980. On the efficiency of a competitive equilibrium in infinite horizon monetary economies. *Review of Economic Studies* 47(4): 797–807.
Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482; 67: 518–522 (1959).
Sheshinski, E., and Y. Weiss. 1981. Uncertainty and optimal social security systems. *Quarterly Journal of Economics* 96(2): 189–206.
Tirole, J. 1985. Asset bubbles and overlapping generations. *Econometrica* 53(5): 1071–1100.

Intergenerational Transmission

Lance Lochner

Abstract

Intergenerational transmission refers to the transfer of individual abilities, traits, behaviours and outcomes from parents to their children. This article analyses the key theoretical and empirical issues in studies of intergenerational transmission of educational attainment, welfare receipt and fertility. Mechanisms that lead to intergenerational transmission of these outcomes are discussed in detail. The role of government policy in affecting intergenerational transmission is also considered.

Keywords

Ability; Altruism; Borrowing; Credit markets; Education; Educational attainment; Educational subsidies; Fertility; Human capital; Inequality; Instrumental variables; Intergenerational correlation; Intergenerational income mobility; Intergenerational transmission; Patience; Preferences; Risk aversion; Skills; Social mobility; Twin studies; Unobserved heterogeneity; Welfare receipt

JEL Classifications

J12; J139

Intergenerational transmission refers to the transfer of individual abilities, traits, behaviours and outcomes from parents to their children. Economists have largely focused on the intergenerational transmission of educational attainment, earnings and income, wealth, fertility decisions and welfare receipt. When intergenerational transmission is strong, children turn out much like their parents, and social mobility is low.

Raw intergenerational correlations in education, earnings, teenage childbearing and welfare receipt in the United States are sizable. Correlations between parents'and children's educational attainment and earnings are both around 0.4. Daughters of teenage or welfare mothers are nearly twice as likely to have a child when they are teenagers compared to daughters of older or non-welfare mothers. Mothers who grew up in a welfare family are four to six times more likely to receive welfare themselves than other mothers.

What gives rise to the intergenerational transmission of these outcomes? Parents may genetically pass on abilities, endowments, or preferences to their children that predispose them to choose actions similar to those they themselves chose. This can generate an intergenerational correlation in outcomes even if there is no actual causal effect of a parent's behaviour or outcome on the child. However, parents' actions themselves may encourage their children to take similar actions. For example, parents' schooling choices may directly impact on their children's decisions to stay in school. Intergenerational transmission incorporates both causal and non-causal channels.

Identifying the mechanisms that lead to the intergenerational transmission of education, earnings, fertility, or welfare receipt is central to understanding the role played by economic conditions or government policies in shaping those

relationships. For example, if differences in earnings or welfare receipt primarily reflect differences in genetically endowed abilities, then policies to expand educational opportunities may have little effect on the intergenerational transmission of earnings and welfare. On the other hand, if ability primarily influences earnings by altering individual access to, or the financial rewards from, schooling, then college subsidies for low-income families should weaken the link between parents' and their children's earnings and welfare receipt.

This article offers detailed analyses of the key theoretical and empirical issues in studies of intergenerational transmission of educational attainment, welfare receipt and fertility. See also intergenerational income mobility for a discussion of earnings and income transmission.

Educational Attainment

The economics literature has emphasized the role of skill and human capital development in analysing intergenerational transmission. To begin, consider an overlapping generations economy that generalizes the model of Becker and Tomes (1986) in which parents choose between investing in their children's human capital, their own current consumption, and borrowing or saving in the form of debts or bequests left for their children. Parents care about their own current consumption, but they also care about the consumption of their children and all future generations. While schooling is costly for parents and children, it raises human capital (or skill) levels, which increases subsequent earnings. Suppose that the production of a child's human capital, H_c, depends positively on parental human capital levels, H_p, the child's 'natural' ability, A_c, and the total years of child schooling, s_c, such that $H_c = h(H_p, A_c, s_c)$. Further, assume that both child ability and parental human capital raise the marginal productivity of schooling (that is, $\partial^2 h/\partial s_c \partial H_p \geq 0$ and $\partial^2 h/\partial s_c \partial A_c \geq 0$). These assumptions imply that: (*a*) for any given level of child investment or schooling, an increase in parental education or child ability produces more child human capital and (*b*) more able children from more educated parents will tend to invest more in their skills through schooling. Finally, assume that the abilities of children and parents are positively (but not perfectly) correlated. That is, bright parents tend to have bright children while dull parents tend to have dull children, but there is, on average, regression to the mean.

If parents are free to leave any amount of bequests/debts to their children, optimal child schooling for each generation will be chosen to maximize discounted earnings less investment costs. In this case, a child's schooling, $s_c = \sigma(H_p, A_c, \pi)$ is an increasing function of his parents' human capital and his own ability, while it is decreasing in the price of schooling, π. Importantly, the optimal schooling level will not depend on parental earnings or wealth, although it may be correlated with both since they depend on parental abilities and human capital. In this simple model, a positive correlation in schooling between parents and children arises for two primary reasons: (1) parental human capital directly raises the productivity of child schooling and (2) abilities are positively correlated across generations and ability raises the productivity of schooling. Economists interested in identifying the 'causal effect' of parental schooling on child schooling attempt to estimate effect (1). This reflects the amount child schooling would increase if policy interventions were to raise parental schooling (and all else were held constant).

Effect (2) depends on the intergenerational transmission of ability. As this is driven by genetics, it reflects the main role played by nature. If a child's human capital depended only on his own ability and schooling (so $\partial h/\partial H_p = 0$ as assumed in Becker and Tomes 1986), only effect (2) would matter, and the intergenerational transmission of educational attainment would be driven by the intergenerational transmission of ability. Even in this case, nurture plays a role in that schooling and other family investments are choices made by families. When schools change their prices (or quality), schooling decisions and the intergenerational transmission of educational attainment are affected.

Imperfect credit markets with limited borrowing opportunities also weaken the link between

ability and schooling for poor families. When poor parents cannot borrow against their own future earnings or leave debts for their children, they may be forced to compromise on both their own consumption and schooling for their children (see, for example, Becker and Tomes 1986; Caucutt and Lochner 2006). Among constrained families, schooling choices depend on family income, I_p, so $s_c = \sigma'(H_p, A_c, \pi, I_p)$ where $\partial\sigma'/\partial I_p \geq 0$. Poorly educated (and, consequently, low-income) parents lucky enough to have bright children may not be able to afford the efficient amount of schooling for them. (This need not be true when parental human capital has a very strong effect on the marginal product of schooling; in this case, poorly educated parents may not want to invest much in their children, even when they are bright.) This implies a strong intergenerational transmission of schooling among the least educated who cannot escape their misfortune. Since more educated and wealthier parents can afford efficient investments in their children, their behaviour is driven by the forces described earlier (that is, $s_c = \sigma(H_p, A_c, \pi)$). That the most disadvantaged underinvest in their children (while the most advantaged do not) when borrowing opportunities are limited implies that policies designed to subsidize the schooling of poor children will help to reduce economic inequality while improving aggregate efficiency.

Most researchers agree that the primary reason many college-age children from poor families do not attend college is that they are ill-prepared and not because they are unable to borrow for college. This raises the question as to whether these youths are ill-prepared because their parents have been unable to borrow the resources needed to prepare them for college in the first place. Direct evidence is scant, but indirect evidence suggests that poor parents sometimes fail to make early educational investments in their children that have substantial long-run payoffs. Cunha et al. (2007), therefore, argue that policies promoting early investments (for example, preschool) in children do not face the same equity-efficiency trade-off that late investments (for example, college or post-school training) do.

The intergenerational transmission of preferences (for example, altruism, patience, or risk aversion) and other causal channels (for example, schooling may stimulate intellectual curiosity that is passed on to children) may also play important roles in the intergenerational transmission of education. While Mulligan (1997) explores the implications of endogeneous altruism, most economists have not incorporated these channels into their theoretical models.

The empirical literature typically considers a linearized version of the schooling decision described earlier:

$$S_{ci} = \alpha S_{pi} + \beta A_{ci} + \gamma I_{pi} + X_i\delta + \varepsilon_i, \qquad (1)$$

where X_i reflects variables that may affect the costs or benefits of schooling (for example, parenting skills, neighbourhood characteristics, school quality, or tuition prices) for child i. With ideal data, estimates from this equation inform us about the schooling choice function. Estimates of α tell us the direct effect of an increase in parental schooling, net of any effects parental schooling has on family income (or neighbourhood and school characteristics included in X). To obtain the total effect of parental education ($\alpha_T = \alpha + \gamma\partial I_p/\partial S_p + \delta\partial X/\partial S_p$), one must incorporate its effects through family income and the X variables. These effects are typically referred to as causal effects, since they measure how much a change in parents' education causes children's education to change. Most empirical studies suggest that the difference between α and α_T is small. See Haveman and Wolfe (1995) or Behrman (1997) for surveys of standard multivariate regression estimates of Eq. (1).

Since data do not typically contain reliable measures of child ability, neighbourhood and school peer quality, or parental skills in bringing up children, most regression-based estimates of Eq. (1) are probably upward biased for α. Researchers have begun to exploit three alternative econometric techniques that aim to reduce or eliminate biases arising from these types of unobserved factors: comparisons of children born of twin mothers or fathers, studies of adopted children, and instrumental variable approaches.

Some researchers have estimated how schooling differences between cousins whose parents are

identical twins depend on the educational differences between their twin parents. This approach assumes that schooling differences among twin mothers or fathers are random rather than the result of different abilities or environments – an assumption often questioned. If the effects of unmeasured ability and parenting skill differences are additively separable from the effects of parental schooling, within-twin-parent estimators remove the effects of genetic differences in parental ability (from the twin parent side of the family) as well as any variation in the twins' parenting skills owing to the similarity of their upbringing – two potential sources of bias. Twin-parent-based estimates generally imply an important role for unobserved ability and parenting skills in determining child schooling levels. Using recent US data on the children of twins, Behrman and Rosenzweig (2002) find that within-twin-parent estimates of the effect of father's schooling are positive and statistically significant, while the estimated effect for mother's schooling is not. That is, differences in schooling between cousins with fathers who are twins are positively correlated with the difference between their fathers' schooling. For cousins with twin mothers, differences in child and differences in mothers' schooling are uncorrelated. (Controlling for differences in spouses' schooling or earnings has little effect on these conclusions.) In explaining the finding that a mothers' schooling does not affect child schooling, the authors argue that more educated mothers spend more time working and may, therefore, spend less time bringing up their children. However, this was not true in the 1970s in the United States (Leibowitz 1974) nor is it true today in rural India (Behrman et al. 1999), where women work little outside the home. This shows that the economic environment plays an important role in determining intergenerational relationships.

A different approach estimates the effects of parents' schooling on adopted children. When the effects of nature and nurture are additively separable and adoptees are randomly assigned to adoptive parents, the estimated effects of the adoptive parents' education on adoptees' schooling eliminates any bias due to the genetic transmission of ability. Under these circumstances, the estimated effects from adoptees provide a measure of the role played by nurture. However, they need not reflect the causal effect of parental education if some unobserved parenting skills are correlated with (but not caused by) parents' educational attainment. Bjorkland et al. (2006) use a unique data-set from Sweden that contains educational attainment for adopted children and *both their biological and their adopting parents*. This enables them to regress adoptee schooling on the schooling of both biological parents, both adoptive parents, and even the interaction of biological and adoptive parents' schooling. While their results suggest important effects of the biological and adoptive father's and biological mother's education on their children, evidence of the adoptive mother's role is mixed. Interestingly, they estimate a positive and significant interaction between the biological and adoptive mother's schooling, suggesting an important nature–nurture complementarity. This interaction raises questions about methods that rely on the assumption that genetic and environmental effects are additively separable (for example, twin-parent studies or other adoptee studies that do not use data on both biological and adoptive parents).

Finally, some recent studies use changes in compulsory schooling laws in the United States and Europe as instrumental variables for changes in parental schooling. The legal changes largely affect the educational outcomes of parents at the low end of the distribution; thus, the studies' findings measure the impacts of increasing schooling among less-educated parents. Furthermore, the laws alter the population distribution of schooling, which may impact marriage markets. As such, they do not necessarily measure the effects of changing a single parent's schooling level. A Norwegian study (Black et al. 2005) estimates little causal effect of parental schooling (except for the mother–son relationship) when using an increase in compulsory schooling as an instrument, but the effects are not very precisely estimated. By contrast, a US study (Oreopoulos

et al. 2006) finds that a mother's and father's education has a significant effect on the probability that a young child is a year behind at school.

To summarize, most researchers conclude that parental education has a causal effect on child education, albeit substantially smaller than raw correlations suggest. While a few recent studies that compare children with twin parents or that focus on adopted children suggest that changes in a mother's education may have very small effects, instrumental variables studies do not confirm this pattern. Adoptee studies suggest that the educational outcomes of biological parents are important even when the child is brought up by others. Thus, the genetic transmission of abilities and preferences plays an important role in intergenerational transmission. Bjorkland et al. (2006) estimate an important interaction between nature and nurture that is often neglected in empirical analyses. Finally, even studies that estimate causal effects do not separately identify the mechanisms by which parents' schooling affects child schooling. We are still left wondering whether schooling changes the preferences or information of parents, or whether it changes the marginal productivity of investing in one's children.

Teenage and Non-marital Fertility and Welfare Receipt

Studies of intergenerational fertility transmission have typically focused on non-marital and teenage births, as these are often associated with a wide range of negative outcomes for mothers and their children. Studies of intergenerational welfare receipt invariably discuss intergenerational patterns for education, earnings, and fertility as well. Economic theories of fertility (for example, Becker 1991) generally say little about intergenerational patterns in childbearing and marital decisions. Formal economic models of intergenerational welfare transmission are also notably absent. Despite a lack of formal theory, social scientists have identified a number of factors that may affect the intergenerational transmission of fertility and welfare outcomes, including intergenerational correlations in cognitive ability, age of puberty, education and earnings. Economists are most interested in causal channels, however. Studies of teenage and non-marital fertility often refer to parental rolemodel effects and the impacts of early/non-marital childbearing on subsequent family structure and economic resources. Studies of intergenerational welfare patterns stress that parental welfare receipt may affect children's views about accepting public transfers, inform children about the welfare system, limit connections in and information about the labour market, and augment family resources.

Empirical researchers primarily aim to estimate the causal effects of parental teenage or out-of-wedlock childbearing and welfare receipt on daughters' choices; however, it is difficult to separate causal effects from other factors that contribute to intergenerational correlations. Analyses typically employ multivariate regression techniques to control for measured family and environmental conditions, but concerns about unobserved heterogeneity plague most studies. Unmarried welfare mothers almost certainly differ from married mothers who are not on welfare, even when current family income and other observable characteristics are the same.

Kahn and Anderson (1992) estimate very different roles of teen motherhood on the fertility decisions of black and white children. They find that teen motherhood largely affects white daughters' marital teen childbearing whereas black daughters' non-marital teen childbearing is most affected. Differences in family background drive much of the intergenerational correlation of teen motherhood for whites but not blacks. Biological links related to the age of puberty play no role in teen fertility for either race. Two more studies (Haveman et al. 2001; Wolfe et al. 2001) separate the effects of the mother's age and her marital status at childbirth on the probability that a daughter has an out-of-wedlock birth as a teenager. The first study finds that mother's age is the more important factor, while the second concludes that

marital status is more important. There is no consensus in the literature as to the relative importance of mother's age or marital status at the time of birth on her daughter's subsequent fertility decisions.

Most empirical studies of intergenerational welfare receipt control for parental income levels (or welfare eligibility), and attempt to estimate how parental welfare acceptance itself affects daughters' future welfare receipt. Some studies use instrumental variables (typically, local unemployment rates or state welfare benefit levels) to further account for unobserved heterogeneity in family tastes or productivity levels (for example, Levine and Zimmerman 1996; Pepper 2000). Gottschalk (1996) exploits the timing of parental welfare receipt (while the daughter lives at home and afterwards) in an attempt to control for unobserved permanent family characteristics. These studies generally conclude that parental welfare receipt increases the daughter's subsequent welfare receipt and childbearing, but much (or even most) of the raw intergenerational correlation is attributed to the correlation in both income and unobserved heterogeneity. Recent studies suggest that there is a small positive causal effect of family income on children's educational outcomes (for example, see Dahl and Lochner 2006); however, most intergenerational welfare studies find that income-enhancing effects from parental welfare payments do not reduce the probability of daughter's welfare receipt enough to offset other direct effects on daughters' tastes or information.

See Also

- ▶ Education Production Functions
- ▶ Family Economics
- ▶ Human Capital
- ▶ Intergenerational Income Mobility

Bibliography

Becker, G.S. 1991. *A treatise on the family*. Cambridge, MA: Harvard University Press.

Becker, G.S., and N. Tomes. 1986. Human capital and the rise and fall of families. *Journal of Labor Economics* 4: S1–39.

Behrman, J. 1997. Mother's schooling and child education: A survey. Working paper No. 97–025, PIER.

Behrman, J., A. Foster, M. Rosenzweig, and P. Vashishtha. 1999. Women's schooling, home teaching, and economic growth. *Journal of Political Economy* 107: 682–714.

Behrman, J., and M. Rosenzweig. 2002. Does increasing women's schooling raise the schooling of the next generation? *American Economic Review* 92: 323–334.

Bjorkland, A., M. Lindahl, and E. Plug. 2006. The origins of intergenerational associations: Lessons from Swedish adoption data. *Quarterly Journal of Economics* 121: 999–1028.

Black, S., P. Devereux, and K. Salvanes. 2005. Why the apple doesn't fall far: Understanding intergenerational transmission of human capital. *American Economic Review* 95: 437–442.

Caucutt, E. and Lochner, L. 2006. Early and late human capital investments, borrowing constraints, and the family. Working paper, University of Western Ontario.

Cunha, F., J.J. Heckman, L. Lochner, and D. Masterov. 2007. Interpreting the evidence on life cycle skill formation. In *Handbook of the economics of education*, ed. E. Hanushek and F. Welch, vol. 1. Amsterdam: North-Holland.

Dahl, G. and Lochner, L. 2006. The impact of family income on child achievement. Working paper No. 11279. Cambridge, MA: NBER.

Gottschalk, P. 1996. Is the correlation in welfare participation across generations spurious? *Journal of Public Economics* 63: 1–25.

Haveman, R., and B. Wolfe. 1995. The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature* 33: 1829–1878.

Haveman, R., B. Wolfe, and K. Pence. 2001. Intergenerational effects of nonmarital and early childbearing. In *Out of wedlock: Causes and consequences of nonmarital fertility*, ed. L. Wu and B. Wolfe. New York: Russell Sage.

Kahn, J., and K. Anderson. 1992. Intergenerational patterns of teenage fertility. *Demography* 29: 39–57.

Leibowitz, A. 1974. Education and home production. *American Economic Review* 64: 243–250.

Levine, P. and Zimmerman, D. 1996. The intergenerational correlation in AFDC participation: Welfare trap or poverty trap? Discussion paper No. 1100–96, Institute for research on poverty.

Mulligan, C. 1997. *Parental priorities and economic inequality*. Chicago: University of Chicago Press.

Oreopoulos, P., M. Page, and A.H. Stevens. 2006. The intergenerational effects of compulsory schooling. *Journal of Labor Economics* 24: 729–760.

Pepper, J. 2000. The intergenerational transmission of welfare receipt: A nonparametric bound analysis. *Review of Economics and Statistics* 82: 472–488.

Wolfe, B., K. Wilson, and R. Haveman. 2001. The role of economic incentives in teenage nonmarital childbearing choices. *Journal of Public Economics* 81: 473–511.

Intergovernmental Grants

Nora E. Gordon

Abstract

Intergovernmental grants are payments from one level of government to another, such as from the federal government to a state government, or from a city to a school district. Theoretically, such grants allow more local choice in public goods provision than purely centralized provision would, while still enabling some redistribution across local jurisdictions. Empirical research on these grants has focused on the extent to which these grants ultimately affect spending by receiving jurisdictions, both on the intended programme area and overall, and on other unintended consequences of the grants.

Keywords

Block grants; Bureaucratic capture; Crowding out; Fiscal federalism; Flypaper effect; Intergovernmental grants; Interjurisdictional spillovers; Matching grants; Public spending; Targeted public spending; Tiebout hypothesis

JEL Classifications

H7

Intergovernmental grants are payments from one level of government to another, such as from the federal government to a state government, or from a city to a school district.

Intergovernmental grants are widely used in the United States across a range of policy functions and are an important tool for redistribution in a federalist context. Under the Tiebout hypothesis, providing public goods locally rather than centrally improves match quality between individual preferences and local provision levels and generates competition in efficiency of public goods provision across communities, limiting bureaucratic capture. In a purely local system, however, any spillovers to public spending across local jurisdictions generate inefficient levels of public spending, and the ability to redistribute is limited to within local borders. Intergovernmental grants provide a mechanism to retain some benefits of local provision, while allowing for more optimal levels of public spending in the presence of interjurisdictional spillovers and increasing the capacity for redistribution.

The economic literature on intergovernmental grants investigates both their fiscal and their non-fiscal effects. Research on the fiscal impact of intergovernmental grants focuses on the extent to which they supplement local revenue formerly dedicated to the programme area, rather than supplanting it. Because intergovernmental grants are used in such a variety of policy functions, they have the capacity – especially if they do not crowd out local revenue – to affect a wide range of non-fiscal outcomes. Before discussing the research on the effects of intergovernmental grants, I briefly discuss the main types of intergovernmental grant structures.

Block Grants and Matching Grants

The most important distinction between block grants and matching grants is that matching grants change the relative prices facing the receiving jurisdiction, making the publicly provided good or service in question relatively cheaper, while block grants provide income but do not change prices. Both types of grant typically are directed to particular agencies or programmes.

Block grants transfer funds from one jurisdiction to another, and are theoretically equivalent to the receiving jurisdiction facing a positive income shock from any source. A conditional block grant requires that the receiving jurisdiction spend at least the grant amount on the governmental activity targeted by the grantor jurisdiction. The extent to which the condition is binding depends on the preferences of the receiving jurisdiction. Despite this constraint, the fungibility of grant income makes it difficult to force receiving jurisdictions to increase spending by the full grant amount. Grantor jurisdictions often attempt to address this issue through 'maintenance of effort' requirements,

by which receiving jurisdictions are required to continue funding the programme to which the grant is dedicated at some set percentage of previous years' levels in order to receive the grant.

When a grantor jurisdiction offers matching grants, it sets a rate at which it will match contributions from the grantee jurisdiction. These rates may vary depending on the level of contributions. Matching grants differ from block grants in fundamentally changing incentives for spending on education by making education spending 'cheaper' than other spending.

Data on Intergovernmental Grants in the United States

The Census of Governments, conducted every five years in years ending in 2 and 7, collects data from states, counties, cities and other municipalities, independent schools districts, and special districts on all revenues and expenditures, including intergovernmental grants. For intergovernmental grants, the Census of Governments details the source of revenue or destination of payments (federal, state, or local) and the policy function to which it is dedicated (for example, health, education, or fire).

Evidence of Fiscal Impacts: the Flypaper Effect

Economic theory predicts that a jurisdiction receiving an intergovernmental lumpsum grant targeted to a particular function of government will view the grant as income and spend it as such, with a fraction going to the targeted function, and the remainder going to other projects or to private consumption through reductions in tax rates. Many empirical studies, however, have observed that the marginal propensity to spend an intergovernmental grant on public expenditures is higher than the marginal propensity to spend other income on public expenditures. Arthur Okun called this phenomenon the flypaper effect, because the money 'sticks where it hits' (see Gramlich 1977). There are three main categories of explanation for these observed effects: (*a*) they are real and reflect the preferences of bureaucrats but not of voters; (*b*) they are real and reflect voters' preferences, but voter preferences may reflect some behavioural anomalies, such as loss aversion and lack of fungibility; (*c*) they are not real, but are generated by econometric misspecification. Hines and Thaler (1995) describe more specific cases within these categories in detail.

Given the current and historical prevalence of such grants, whether they ultimately supplement, or 'stick to', local spending is, unsurprisingly, the subject of a lengthy empirical literature, of which Hines and Thaler provide an excellent review. Studies included typically find that intergovernmental grants increase expenditures on the targeted programme by 25 to 100 per cent of the grant amount, with most estimates clustered at the high end of the range. This is much more than the receiving government's estimated propensity to spend on public programmes out of regular income (here Hines and Thaler estimate that only five to ten per cent of new non-grant income would be spent on public programmes), corresponding to a strong flypaper effect. One of the most convincing studies in their review is that of Ladd (1992), which shows that plausibly exogenous increases in state tax bases (stemming from the fact that some states link their tax base definition to the federal one, and exploiting changes in the federal income tax base following the Tax Reform Act of 1986) generate increases of about 40 per cent in state revenue. Many other studies simply correlate intergovernmental grants with spending, often in a cross-sectional context, without regard to potential bias from the fact that the same factors which make some jurisdictions receive more intergovernmental payments in a particular policy area may also make them have higher demand for public spending in that area.

Several recent additions to this literature have focused more explicitly on isolating exogenous variation in grant levels, and in doing so have yielded much less 'sticky' results. Knight (2002) accounts for political endogeneity in the amount of federal highway aid received by states by exploiting variation in legislative bargaining power due to seniority of state representatives in the US House. His technique reveals significant

crowd-out of states' own support of their highway programmes.

A number of recent papers focus on the heterogeneity of flypaper effects. Gordon (2004) shows that governments receiving intergovernmental grants may need time to adjust other revenue sources in response. Federal Title I grants to school districts for compensatory education, based largely on child poverty counts, appeared to stick completely to school spending in the first year following a shock to grant amount after the release of new census poverty data. Three years after the shock, however, there appears to be no effect on spending. Baicker and Staiger (2005) highlight the importance of institutional factors in determining how much receiving jurisdictions are capable of crowding out. In examining state responses to federal Medicaid Disproportionate Share Hospital (DSH) grants, they find that states which allow different levels of government to transfer funds directly between one another crowded out about half the federal grants. In states without this institutional capacity, the DSH funds were much stickier. Strumpf (1998) shows that the share of local spending on administrative overhead (a proxy for bureaucratic power) predicts the extent to which intergovernmental payments stick to local budgets, supporting a bureaucratic capture explanation of the flypaper effect.

Evidence of Non-fiscal Impacts

Intergovernmental grants have a wide range of effects, intended and unintended, on non-fiscal outcomes. The intended effects of intergovernmental grants may be due to the productive use of the grant. For example, Baicker and Staiger (2005) go on to show that federal DSH grants have significant impacts on mortality, despite the substantial crowd-out observed. Their findings suggest that the effects on mortality are due to the sticky part of the grant, which improves quality of hospital care. More often, studies evaluate the effect of the total intergovernmental grant amount rather than the effective or sticky grant amount on the outcome targeted by the grant. Such studies may conclude that public spending in that area is not effective, when in fact other revenue was crowded out so that total public spending in that area did not rise.

Jurisdictions making intergovernmental grants may do so to create incentives for the receiving governments that differ from simply spending the payment as designated. For example, Title I of the Elementary and Secondary Education Act of 1965 strengthened incentives for school districts to desegregate in compliance with the Civil Rights Act of 1964, and school districts responded accordingly (Cascio et al. 2005), though Title I funded compensatory education activities rather than desegregation-related costs. The current incarnation of this programme, the No Child Left Behind Act of 2001, similarly uses the threat of losing compensatory education funds as an incentive for schools to meet criteria for academic achievement growth benchmarks.

Finally, intergovernmental grants may create incentives that generate consequences unintended by the granting jurisdiction. For example, Cullen (2003) attributes 40 per cent of the significant rise in the special education classification of Texas public (government) school students from 1991 to 1996 to increased payments from the state to districts on a per-classified-student basis.

See Also

- ▶ Fiscal Federalism
- ▶ Tiebout Hypothesis

Bibliography

Baicker, K., and D. Staiger. 2005. Fiscal shenanigans, targeted federal health care funds, and patient mortality. *Quarterly Journal of Economics* 120: 345–386.

Bailey, S.J., and S. Connolly. 1998. The flypaper effect: Identifying areas for future research. *Public Choice* 95: 335–361.

Cascio, E., N. Gordon, E. Lewis and S. Reber. 2005. Financial incentives and the desegregation of Southern schools. Working paper.

Cullen, J.B. 2003. The impact of fiscal incentives on student disability rates. *Journal of Public Economics* 87: 1557–1589.

Gramlich, E.M. 1977. Intergovernmental grants: A review of the empirical literature. In *The political economy of*

fiscal federalism, ed. W.E. Oates. Lexington: Heath Publishers.
Gordon, N. 2004. Do federal funds boost school spending? Evidence from Title I. *Journal of Public Economics* 88: 1771–1792.
Gruber, J. 2005. *Public finance and public policy*. New York: Worth Publishers.
Hines, J.R., and R.H. Thaler. 1995. Anomalies: The flypaper effect. *Journal of Economic Perspectives* 9(4): 217–226.
Knight, B. 2002. Endogenous federal grants and crowd-out of state government spending: Theory and evidence from the federal highway aid program. *American Economic Review* 92: 71–92.
Ladd, H. 1992. State responses to the TRA86 revenue windfalls: A new test of the flypaper effect. *Journal of Policy Analysis and Management* 12: 82–103.
Oates, W.E. 1999. An essay on fiscal federalism. *Journal of Economic Literature* 37: 1120–1149.
Strumpf, K.S. 1998. A predictive index for the flypaper effect. *Journal of Public Economics* 69: 389–412.
Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Internal Economies

G. Becattini

Abstract

Marshall introduced the idea of 'internal' economies, which accompany the growth of the 'individual representative firm', as opposed to the 'external' economies accompanying the growth of 'a national or a local industry'. In principle the pursuit of internal economies would lead to a world composed of firms each one producing a great share of a very small range of commodities. But while Marshall shared the classical view of an increasing average size of the business unit, he put it into a dynamical and historical context. Tendencies and countertendencies may result in different outcomes in terms of market structures.

Keywords

Concentration; External economies; Internal economies; Market structure; Marshall, A.; Mechanization

JEL Classifications

D2

The expression 'internal economies' is considered here solely in terms of Alfred Marshall's own formulation, which is quite different from current terminology referring to internal economies of scale. Modern terminology refers to a reduction in the average cost of production of a well-specified commodity in relation to increases in the quantity produced, assuming, for every given quantity produced, the most appropriate utilization of the optimum productive plant. Marshall's concept of internal economies is analytically looser than this, but richer in empirical content and, possibly, in philosophical insight.

The twin terms, 'internal' and 'external' economies (and diseconomies) were first used by Marshall 'for indicating the fundamental distinction between the "internal" economies and wastes which come with an increase in the size of the individual representative firm; and those "external" economies and wastes which come with an increase in the aggregate volume of a national or a local industry' (Marshall 1890, vol. 2, p. 347).

The main aim of the distinction was to help the applied economist in his attempts to disentangle the intricacies of contemporary socio-economic reality, rather than to provide an integral part of a formal theory of the relative values of the commodities; this is shown clearly enough by the ambiguous references to 'national or local industry' and by the use of such a fuzzy concept as the 'individual representative firm'. We must add that many times Marshall conveys the impression of confining external economies in the straitjacket of a single-product, homogeneous industry. Moreover, we must bear in mind, as Loasby aptly pointed out, that Marshall 'made no clear distinction between the theory of value and the theory of growth' (Loasby 1978, p. 1, n. 1).

This vein of Marshallian thought derives from three sources: his vast and detailed knowledge of the literature on contemporary British and American industry; his own ruminations on the Smith–Babbage arguments on the division of labour and the internal organization of the firm,

and, finally, his own early studies of mental science.

There is a passage in the *Principles* that contains the kernel of the Marshallian ideas on the internal growth of the firm. 'Practice makes perfect', starts Marshall, taking up the well-known Smithian theme:

> physiology, [he continues], in some measure explains this fact. For it gives reasons for believing that the change is due to the gradual growth of new habits of more or less reflex or automatic action. Perfectly reflex actions . . . are performed by the responsibility of the local nerve centre without any reference to the supreme central authority of the thinking power, . . . But all deliberate movements require the attention of the chief central authority: it receives information from the nerve centre or local authorities and perhaps in some cases direct from the sentient nerves, and sends back detailed and complex instructions to the local authorities, or in some cases direct to muscular nerves, and so co-ordinates their action as to bring about the required results. (1890, vol. 1, pp. 250–1)

This quotation helps us put together the scattered pieces of the Marshallian theory of the growth of the firm under competitive conditions.

Under the spell of all the usual drives of the human mind (money-making propensity, 'instinct of the chase, desire for fame', and so on), a business unit, working in a competitive context, is subject to a continual pressure to rationalize its most typical recurring operations and the tools used. So we have, simultaneously, both the development of 'skills' (a 'sort of capital of nerve force'), allowing the saving of time and of physical and, above all, nervous energies, and a rationalization of the process and the tools used. Alert to the danger of sliding to an abstract conception of the industrial process, Marshall makes room for historical and geographical peculiarities of the 'skilling' and 'rationalizing' processes.

But there comes a point 'when the action has thus been reduced to routine [that] it has nearly arrived at the stage at which it can be taken over by machinery' (1890, vol. 1, p. 254) At this point it is very probable that someone will invest the money and the inventive power required for the realization of the appropriate appliance.

When a machine is introduced into a manufacturing firm, its product becomes more uniformly specified and a cumulative process of mechanization and standardization can start. Marshall speaks of 'a great architectonic principle' according to which

> a well-driven machine tool could become the parent of new machine work more exact than itself . . . and so on . . . By successive steps larger and more delicate work is thrown upon the apparatus . . . at last it becomes . . . a thinking acting on hints given from within.. .. When all is in order, the machine is nearly self-sufficient. (1890, vol. 1, pp. 206–7)

The gradual introduction of specialized machinery results in more time and more nervous energies being made free at the hierarchical summit of the firm, in such a way that the entrepreneur can devote more of his time and energies to the 'broadest and most fundamental problems of his trade' (1890, vol. 1, p. 284), that is, to the collection and evaluation of information about general market trends and technological and organizational innovations.

The growing of a business unit above the other units of an industry gives to it the opportunity of taking advantage of a better allocation of skills, of getting hold of 'big brains', of introducing innovations out of reach of the others, of obtaining better terms in buying, selling and borrowing. And consequently, in the words of Marshall: 'lowers the price at which he can afford to sell' (1890, vol. 1, p. 315).

The basic constraint to the development of the individual firm lies in the conflict between the urge of the entrepreneur to decipher the environmental conditions of growth and the organizational requirements of the productive process. From this second viewpoint the best results can be attained by concentrating the entrepreneur's efforts on a narrow range of tasks. The simpler the work of direction, the larger the volume of output which can be efficiently controlled by a single mind, the greater the scope for the introduction of machines and uniform continuous processes. It would seem that the combined effect of these constraints would be a world composed of firms each one producing a great share of a very small range of commodities.

But this outcome would be apparently self-destroying for a world of competitive (albeit

imperfectly) firms, like the Marshallian one. Marshall's answer to this challenge is both complex and stimulating. First of all, to make use of all the possible internal economies, a certain amount of individual volition is needed. The entrepreneur 'works hard and lives sparely . . . subordinates trust him and he trusts them . . . every improved process is quickly adopted .. .'. If this behaviour 'could endure for a hundred years, he and one or two others like him would divide between them the whole of that branch of industry in which he is engaged'. But life is short and those who follow are not always fit to take over the task. The firms of many industries, at least before 'the great recent development of vast joint-stock companies, which often stagnate but do not readily die', like the trees of the forest 'gradually lose vitality and one after another . . . give place to others'.

We must also remember that 'many of the lines of division between the trades which are nominally distinct are becoming narrower and less difficult to be passed . . . A watch factory with those who worked in it could be converted without any overwhelming loss into a sewing-machine factory' (1890, vol. 1, pp. 258–9). This continual trespassing of 'industrial' borderlines systematically frustrates the inner tendencies towards concentration and monopolization.

It must also be taken into account that the continual formation of economies, external to the single firm but internal, either to an industry or to some group of industries, in that they apply even to the smallest firms, systematically erodes part of the advantage of the bigger businesses. A particularly relevant example of this is provided by the case of a localized population (the Marshallian 'industrial district') of medium-small sized firms, which, grouping together and specializing in various stages of the production process, achieve many of the large-scale economies typical of the giant firms.

Marshall shares the classical view of an increasing average size of the business unit, but he is very careful to put it into a dynamical and historical context.

Tendencies and countertendencies may result in different outcomes in terms of market structures. What is necessary for the process to be self-perpetuating is that the system should reproduce the complex of motivations which, given the structural characteristics of the industrial field, nourish the basic tendency of man towards liberation from purely mechanical tasks. In the words of R.A. Jenner: 'external and internal economies thus form counterbalanced forces of competition around which the disturbing thrusts of evolutionary change are held in control' (Jenner 1964, p. 311).

See Also

- ▶ External Economies
- ▶ Returns to Scale

Bibliography

Jenner, R.A. 1964. The dynamic factor in Marshall's economic system. *Western Economic Journal* 3 (1): 21–38.

Loasby, B.J. 1978. Whatever happened to Marshall's theory of value? *Scottish Journal of Political Economy* 1: 1–12.

Marshall, A. 1890. *Principles of economics*, ed. C.-W. Guillebaud. 9th (Variorum) edn, 2 vols. London: Macmillan, 1961.

Marshall, A. 1919. *Industry and trade*. London: Macmillan.

Internal Migration

James R. Walker

Abstract

Migration is a shared topic within social sciences attracting interest from members of all sub-disciplines. This attention reflects both the importance of the flows and the complexity of the behaviour. This article presents a short overview of the basic theoretical perspectives on individual migration decision making, and it considers empirical challenges to bringing these models to the data.

Keywords
Family decision making; Human capital; Internal migration; Labour market search; Learning; Social networks

JEL Classifications
J60

Migration is a shared topic within social sciences attracting interest from members of all subdisciplines. This attention reflects both the importance of the flows and the complexity of the behaviour. This article presents a short introduction to economic analyses of internal migration.

Theory

Since the seminal work of Sjaastad (1962), economists have recognized that migration is a form of human capital. In the simplest model of wealth maximization the fixed costs of moving are balanced against the net present value of earnings streams available in the alternative location. This framework explains why, as was first noted by Ravenstein (1885), migration is an activity primarily of the young. The young are most likely to move, according to the human capital perspective, for three related but distinct reasons. First, they should move to take advantage of economic opportunities as soon as they are independent economic actors. Second, the young have a longer horizon over which to amortize the fixed cost of migration; hence, relatively small gains in earnings may tip the scales in favour of moving. And third, the young have fewer location-specific investments that serve to tie them to the current location (such as children).

Since its publication, Sjaastad's framework has been extended in a variety of ways (see Greenwood (1997), for a useful summary). Wealth maximization, as a motivation for migration, has given way to utility maximization, with uncertainty, information and local amenities given special attention. Perhaps the richest set of behavioural models appears in the development literature, where models of family behaviour incorporating notions of risk sharing, intergenerational transfers, and household bargaining have been developed. (See Lucas (1997), for a comprehensive summary, and Stark (1991), for several case studies which tailor the model to a particular context or issue.)

Nearly all the research on migration adopts a static framework, usually within a binary mover–stay decision framework. A classic example is Mincer's paper 'Family Migration Decisions' (1978), which is an early contribution to the now popular area of decision making in multiperson households. Mincer assumes wealth maximization and that spouses have separate preferences and different opportunities across locations. His basic insight is that the location of an individual's maximum may not coincide with the location of joint maximum. Indeed, the location of the joint maximum may not coincide with the location of the individual maximum of either spouse. This gives rise to the concepts of 'tied movers' and 'tied stayers' and sharp predictions on who should remain married and who should separate. One of his interesting predictions is that the incidence of migration should increase soon after a divorce or separation as the now independent individuals move from their 'tied' locations. Mincer also predicts that these forces become stronger as women's labour force participation and earnings increase.

The limitations of a static framework are also evident in Mincer's paper. (It is noteworthy that to date no one has extended Mincer's work in a meaningful way.) Mincer presumes marriage and does not investigate who marries whom (forward-looking agents may consider the possible consequences of different spatial opportunities before consummating the match). And, restricted to a single period, the analysis cannot investigate the timing or temporal sequence of separation and migration. Indeed, to study temporal linkages of migration and other important lifecycle choices (such as marriage or retirement) requires a dynamic framework. And, as illustrated above, a static framework begs the question as to the nature of initial equilibrium. Allowing households to make multiple migration decisions substantially increases the model's complexity. Now the model

must determine *where* and *when* to move. Moreover, prior moves influence subsequent opportunities, giving these models their own natural dynamics.

Empirical Implementation

One of the first empirical regularities gleaned from individual migration histories is their diversity (DaVanzo and Morrison (1981), is an early contribution). The richness of the life histories appears in the diverse terminology describing the types of moves. Concepts such as 'repeat' (an individual's second or higher order move), 'onward' (a move to a new location – all first moves are 'onward') and 'return' (movement back to a previous location, most commonly the individual's childhood location or self- identified 'home') appear. At an aggregate level, notions of 'circular' and 'chain' migration are commonly used.

Data from the US National Longitudinal Survey of Youth, 1979 Cohort (NLSY79) (US Department of Labor 2006), can be used to give an estimate of the magnitude of these flows. Data through 1994 when cohort members were in their mid-30s show that roughly 80 per cent of the cross-sectional sample had never moved out of their childhood state of residence, considered as their 'home' state. Of the 20 per cent of movers, more than half move again, and of the repeat movers, approximately 55 per cent ever return to their home state. Interestingly, few differences appear by gender, but the proportion of movers is U-shaped in completed education – individuals with a high school degree move the least, whereas those with some college or less than a high-school education are more likely to move. (Long (1988, 1991, 1992), and the numerous reports of William Frey at the University of Michigan and the Brookings Institute are among the best sources of descriptive evidence on internal migration flows in the United States.)

To study the influence of labour market opportunities on migration, we would like to define locations corresponding to distinct local labour markets. Within the United States, if we define local labour markets crudely as equivalent to counties, the model admits a choice set of approximately 3100 elements. (Models of residential choice and occupation are sometimes said to be isomorphic. From an empirical standpoint they are not. Models of occupation choice typically have relatively few alternatives, say five or ten, and educational or experience requirements or other characteristics offer a natural ordering to the occupational alternatives. See Neal (1999), for a recent contribution.) Consequently, there is a fundamental trade-off between the economic definition of locations and statistical measures available. For this reason many studies of internal migration use the decennial Census. Yet the decennial Census has its limitations, most importantly that virtually all individual and household characteristics are measured as of the date of the census. Census data offer detailed descriptive summaries of migration flows over narrow geographic regions, but, with no measures of pre-migration characteristics, are of problematic use for unravelling cause and effect.

Extending the analysis to panel data and multiple decision periods makes greater demands on the data and the analysis. Opportunities must be measured for each period of time, and some decision must be made on the persistence of economic opportunities. As in models of job search, the analyst must decide whether 'recall' is available: do migrants have the ability to remember and possibly return to a previous wage offer? If so, the size of the state space within the dynamic program formulation increases exponentially with the agent's memory length. And an important empirical challenge for dynamic analyses is sample attrition, as not being able to locate a respondent who has moved is one of the reasons for not securing an interview. (Survey organizations quickly developed expertise in locating respondents in the early years of the large-scale surveys such as the Panel Studies of Income Dynamics and the National Longitudinal Surveys. Most commonly, respondents are not interviewed because they refuse, not because they could not be located by the survey organization. See Olsen and Reagan (2000), for detailed information on the experience for the NLSY79.)

Nevertheless, Bellman's principle can be usefully applied to represent the decision problem of the individual (or household). Kennan and Walker (2005) adopt a dynamic programming approach for analysing the migration histories within the National Longitudinal Survey of Youth, 1979 Cohort. We find that earnings are an important (economic and statistical) determinant of migration flows, and their inclusion significantly improves the model's fit to the migration flows within the NLSY79. Respondents in the NLSY79 are more likely to leave a poor local labour market but do not necessarily move to 'the best' labour market as predicted by the model. Our findings are consistent with the interpretation that economic factors are an important determinant of migration, but not the *only* factor.

Research Frontiers

Research on internal migration flows remains on the frontier. Models and analysis of life-cycle migration are still in their infancy, with plenty of room for growth. An important challenge to decision theorists is developing frameworks for understanding return migration – why it is optimal to leave and return home. (Learning or more generally the resolution of uncertainty will be part of the explanation. For an early investigation based on learning, see Pessino (1991).)

Investigating the timing and the relationship between migration and other lifecycle choices such as marriage and retirement is another likely active research area. Certainly the migration behaviour by baby boomers will be of increasing interest to federal and local policymakers.

There is broad consensus that economic and family factors are the primary determinants of internal migration flows. Yet no analysis satisfactorily combines both factors. The barriers to doing so are more empirical than conceptual. The set of family members who may potentially influence migration choice is large, with no consensus as to which relationships must be surveyed. Except for the central role of parents and children, there is little additional information to guide one's choice. Obtaining information on the spatial distribution of family members and their avenues of influence (for example, income pooling or information sharing) is time consuming and thus costly. The influence of broader social networks need also be considered. The Great Black Migration within the United States during the first half of the 20th century illustrates the importance of social factors for migration streams (Lemann's (1991) classic *The Promised Land: The Great Black Migration and How it Changed America* offers an engaging account of the family, social and economic factors that stimulated these flows). Migration research may offer another avenue to explore the influence of social interactions, and perhaps provide stronger ties among social science disciplines.

See Also

- ▶ Collective Models of the Household
- ▶ Family Decision Making
- ▶ Human Capital
- ▶ Labour Supply

Bibliography

DaVanzo, J.S., and P.A. Morrison. 1981. Return and other sequences of migration in the United States. *Demography* 18: 85–101.

Frey, W.H. 2006. US demographic publications: Research reports by William H. Frey. Online. Available at http://www.frey-demographer.org/reports.html. Accessed 21 Nov 2006.

Greenwood, M. 1997. Internal migration in developed countries. In *Handbook of population and family economics*, ed. M. Rosenzweig and O. Stark, vol. 1b. Amsterdam: North-Holland.

Kennan, J. and J.R. Walker. 2005. The effect of expected income on individual migration decisions. Unpublished manuscript, University of Wisconsin-Madison.

Lemann, N. 1991. *The promised land: The great black migration and how it changed America*. New York: Vintage Books.

Long, L.H. 1988. *Migration and residential mobility*. New York: Russell Sage.

Long, L.H. 1991. Residential mobility differences among developed countries. *International Regional Science Review* 14: 133–147.

Long, L.H. 1992. Changing residence: Comparative perspectives on its relationship to age, sex, and marital status. *Population Studies* 46: 141–158.

Lucas, R.E.B. 1997. Internal migration in developing countries. In *Handbook of population and family economics*, ed. M. Rosenzweig and O. Stark, vol. 1b. Amsterdam: North-Holland.
Mincer, J. 1978. Family migration decisions. *Journal of Political Economy* 86: 749–773.
Neal, D. 1999. The complexity of job mobility of young men. *Journal of Labor Economics* 17: 237–261.
Olsen, R., and P. Reagan. 2000. You can go home again: Evidence from longitudinal data. *Demography* 37: 339–350.
Pessino, C. 1991. Sequential migration theory and evidence from Peru. *Journal of Development Economics* 36: 55–87.
Ravenstein, E.G. 1885. The laws of migration. *Journal of the Statistical Society of London* 48: 167–235.
Sjaastad, L. 1962. The costs and returns of human migration. *Journal of Political Economy* 70: 80–89.
Stark, O. 1991. *The migration of labor*. London: Basil Blackwell.
US Department of Labor. 2006. National Longitudinal Survey of Youth 1979. Online. Available at http://www.bls.gov/nls/nlsy79.htm. Accessed 30 Nov 2006.

Internal Rate of Return

Harald Hagemann

JEL Classifications
E2

The internal rate of return of an investment project is that discount rate or rate of interest Y which makes the stream of net returns x_t associated with the project equal to a present value of zero. It is the solution for i in the following equation in which indicates the physical lifetime of the investment project.

$$C(0,\theta) = \sum_{t=0}^{\theta} x_t(1+i)^{-t} = 0.$$

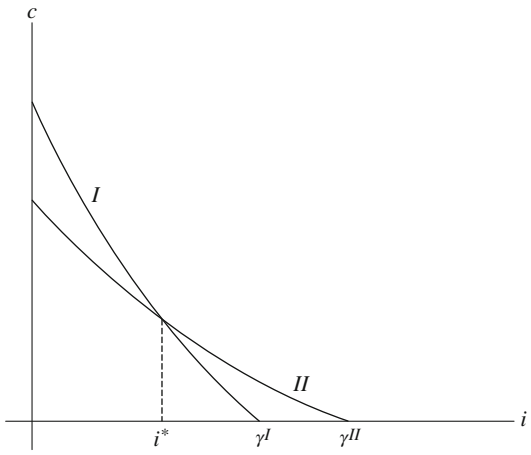
The internal rate of return is compared with the market rate of interest in order to determine whether a proposed project should be undertaken or not.

Among the criteria to be used in determining the profitability of an investment project two others are frequently considered. Whereas the payout-period criterion is a crude rule of thumb which for much of the time ignores pattern of receipts, the net present value criterion is the most relevant 'rule' for optimal investment behaviour. If the present value (using the market rate of interest as the rate of discount) of a project's expected earnings is greater than its cost (including discounted future operating and maintenance costs), that is, if the net present value is positive, the investment project is potentially worth undertaking.

Whereas the net-present-value rule and the internal-rate-of-return rule lead to identical results in the two-period case and in the perpetuity case (which in essence is only a variant of the former), the two criteria may lead to different results in the multiperiod case. Figure 1 illustrates such a case in which the choice between two alternative investment options will lead to identical results for $i > i^*$ whereas the two criteria lead to different results for market rates of interest smaller than the cross-over rate i^* where the present value of i is higher while II has the higher internal rate of return. The failure of the internal rate of return criterion is the consequence of the implicit assumption that all intermediate receipts, positive or negative, are treated as if they could be compounded at the '*internal*' rate of return itself whereas the only appropriate *external* discounting rate is the market rate of interest (reinvestment problem).

When the investment projects are independent and with a perfect capital market (in which the lending and borrowing rates of interest are identical) the net present value is, in general, the only universally correct criterion of appraising investment projects (see Hirshleifer 1958, 1970, ch. 3). For the multiperiod case the internal- rate-of-return rule is not generally correct. Furthermore, there may be *multiple rates of return* that will equate the present value of a project to zero. A necessary condition for non-uniqueness of the internal rate of return is that there be more than one change of sign in the stream of receipts over the lifetime of a project.

The controversy about the multiplicity of the internal rate of return in the late 1950s led to the development of the *truncation theorem.* This



Internal Rate of Return, Fig. 1

theorem turns out to be important for the general problem of choosing the optimal investment period (for a historical survey of truncation theorems see Matsuda and Okishio 1977). In 1969 Arrow and Levhari presented a new version of the truncation theorem which contrasted sharply with the other economists' method of choosing a truncation period so as to maximize the internal rate of return. They rightly pointed out that this criterion would not be adequate for the choice of the truncation period. Instead they advocated the maximization of the present value of the investment project as the proper criterion. It was demonstrated that the possibility of truncating investment projects at any age different from their physical lifetimes and at no extra costs leads to the following results:

(1) The maximized present value of the project is a monotonically decreasing function of the rate of interest. A corollary of this is that the internal rate of return is always unique.
(2) A rise (fall) in the rate of interest will always lower (raise) the present value of the remaining future net returns at all stages of the production process.

Consequently the optimal economic lifetime, too, is a monotonically decreasing function of the rate of interest.

Flemming and Wright (1971) dropped the assumption of a constant rate of discount per unit of time and tried to generalize the theorem to the case of different interest rates over time, a case where the deficiency of the internal-rate-of-return rule is most obvious. However, the 'generalization' does not take us very far because the calculation would require perfect foresight of future rates. The authors emphasize that a 'slight relaxation' of this assumption is allowed because 'a change in expectation which causes' all rates 'to be revised in the same "direction" will alter the present values of all costlessly terminable projects ... in a common direction' (Flemming and Wright 1971, p. 262). But even this proposition holds, in general, only when the change takes place uniformly, so that there is no change in the weights of the time pattern of the stream of net returns.

More interesting is the discussion of the impact of a consequence stream, that is, costs and benefits following from truncation. Whereas a positive *scrap value* can easily be incorporated the range of validity of the truncation theorem is severely limited in the case of *shut down costs.* Shut down costs can occur before and after truncation. Sen (1975) has shown that in the general case of a consequence stream following from truncation only minimal sufficiency conditions can be formulated: non-negative consequence sums (NCS) and non-negative consequence remainders (NCR), that is, the present value of the consequence stream for each t before and after the actual point θ of truncation has to be non-negative. Neither NCS nor NCR requires the present value of the consequence stream at θ to be non-negative, that is, a negative present value of the remaining process does not endanger the monotonicity result. But the conditions are very restrictive, because NCR is violated if the last item or the discounted value of the tail of the consequence stream is negative. This may be the case because of for example, redundancy payments, environmental protection or shut down costs of a nuclear power station.

The truncation theorem was originally developed in a *partial framework.* Nevertheless, Hicks (1973) and Nuti (1973) considered it applicable in a *general framework.* However, Eatwell's (1975)

criticism of these authors has clarified that important propositions of the theorem do not carry over to the general framework (see also Hagemann and Pfister 1978). At the partial level all prices in the economy are taken as given, that is, the individual's stream of net returns is considered not to be affected by changes in the discount rate. This assumption is impermissible when considering investment processes for society as a whole. At the general level the rate of profit is represented by the internal rate of return of the process as a whole for a given real wage. A variation of the discount factor, that is, the profit rate implies an opposite variation of the real wage rate. Because the present value of the whole process is both maximum and zero in competitive equilibrium the slope of the wage–profit curve is negative throughout. This is the only result one can draw under the conditions of the truncation theorem in the general setting. Neither the inverse relationship between the present value of the rest of the process and the rate of profit nor that between the optimal economic process length and the rate of profit invariably hold.

Furthermore, the analysis raises serious doubts as to the existence of an inverse monotonic relationship between interest and investment. The implication for Keynes's concept of the 'marginal efficiency of capital' is close at hand. As is well known, Keynes considered his concept 'identical' with Fisher's definition of the 'rate of return over cost' and stressed that there is no material difference 'between my schedule of the marginal efficiency of capital or investment demand-schedule and the demand curve for capital contemplated by some of the classical writers' (Keynes 1936, pp. 140 and 178). To be sure, there are passages which indicate that Fisher was aware of the fact that prices and therefore not only the present values of the streams of net receipts but the net receipts themselves vary with variations in the rate of interest (see especially the 'more intricate than important' complication discussed in Fisher 1930, pp. 170–71). However, the fixed-price assumption he commonly referred to implies a partial framework where the relationship between interest rates and prices is eliminated. It is therefore impossible to construct a demand curve for investment on the basis of a *ceteris paribus* clause for prices simply by variations of the rate of interest. An inverse macroeconomic relation between interest and investment cannot be derived from monotonicity results reached in a microeconomic framework. The difficulties encountered by Fisher and Keynes are discussed by Alchian and Garegnani from different points of view. Alchian (1955, p. 942) stresses that 'a schedule of investment demand at different market rates of interest requires that one compute the internal rates of return in terms of the prices that would prevail at each potential market rate of interest'. Garegnani (1978–9) brings into focus the problems involved in Keynes's concept of the schedule of the marginal efficiency of capital.

The return of the same truncation period and reswitching of techniques are closely linked phenomena occurring in a general framework. Some authors have tried to draw another analogy between the reswitching problem and the well-known possibility of the existence of multiple rates of return. Apparently the intention was to play down the importance of reswitching. This is reflected by the proposition that 'there is no new thing under the sun' (Bruno et al. 1966, p. 553). However, multiple internal rates of return are a phenomenon related to the partial framework from which a generalization to the general level is not admissible. Truncation ensures the uniqueness of the internal rate of return but cannot rule out reswitching. Therefore, an analogy between the two phenomena does not exist.

See Also

► Investment Decision Criteria

Bibliography

Alchian, A.A. 1955. The rate of interest, Fisher's rate of return over costs and Keynes' internal rate of return. *American Economic Review* 45: 938–943.

Arrow, K.J., and D. Levhari. 1969. Uniqueness of the internal rate of return with variable life of investment. *Economic Journal* 79: 560–566.

Bruno, M., E. Burmeister, and E. Sheshinski. 1966. The nature and implications of the reswitching of techniques. *Quarterly Journal of Economics* 80: 526–553.

Eatwell, J. 1975. A note on the truncation theorem. *Kyklos* 28(4): 870–875.
Fisher, I. 1930. *The theory of interest*. New York: Macmillan.
Flemming, J.S., and J.F. Wright. 1971. Uniqueness of the internal rate of return: A generalisation. *Economic Journal* 81: 256–263.
Garegnani, P. 1978–9. Notes on consumption, investment and effective demand, I and II. *Cambridge Journal of Economics* Pt. I, 2(4): 335–353, December 1978; Pt II, 3(1): 63–82, March 1979.
Hagemann, H., and J. Pfister 1978. Zur Relevanz des Truncation-Theorems in partialanalytischer und totalanalytischer Sicht. *Jahrbücher für Nationalökonomie und Statistik* 193(4): 359–379.
Hicks, J. 1973. *Capital and time. A neo-Austrian theory.* Oxford: Clarendon Press.
Hirshleifer, J. 1958. On the theory of optimal investment decision. *Journal of Political Economy* 66: 329–352.
Hirshleifer, J. 1970. *Investment, interest, and capital*. Englewood Cliffs: Prentice-Hall.
Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan.
Matsuda, K., and N. Okishio. 1977. Theorems of investment truncation. *Annals of the School of Business Administration*, Kobe University 73–90.
Nuti, D.M. 1973. On the truncation of production flows. *Kyklos* 26(3): 485–496.
Sen, A. 1975. Minimal conditions for monotonicity of capital value. *Journal of Economic Theory* 11(3): 340–355.

International Capital Flows

Shang-Jin Wei

Abstract

Cross-border capital flows may be regarded as either too small (known as the Lucas paradox) or too big (against the Samuelson theorem of factor price equalization). The resolution to the conflicting views may require thinking out of the neoclassical box. In theory, international capital flows can promote economic growth, but the data do not reveal a strong, robust, and causal effect, particularly for developing countries. The theoretical results and the empirical patterns can be reconciled through either a composition effect or a threshold effect. Some emerging evidence suggests that the two effects are related.

Keywords

Composition hypothesis on capital flows; Corruption; Economic growth; Factor price equalization theorem; Financial globalization; Foreign aid; Foreign debt; Foreign direct investment; Heckscher–Ohlin–Samuelson model; Institutional quality; International capital asset pricing model (ICAPM); International capital flows; Lucas paradox; Portfolio equity flows; Threshold hypothesis on capital flows; Total factor productivity; Trade costs

JEL Classifications

F2; F3

Cross-border capital flows worldwide have risen substantially since the mid-1970s, from US$1.2 trillion in 1980 to $5.8 trillion in 2004. The pace of the growth (at an average annual rate of 6.6 per cent) surpasses by a big margin those of the world GDP (at 1.7 per cent per annum) and the world exports (at 3.1 per cent per annum). Developed economies are the most important source countries, accounting for 92 per cent of the aggregate outward capital flows in 2004. They are also the most important recipients, accounting for 91 per cent of the aggregate inward capital flows in 2004. A small number of developing countries – commonly known as emerging market economies – receive the lion's share (nearly 70 per cent) of the remaining international capital flows in 2004. More than 130 other developing economies are more or less bypassed by the surge in the capital flows. (For these calculations, developed countries consist of the following 25 countries: Australia, Austria, Belgium, Canada, Cyprus, Denmark, Euro Area, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Japan, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, the United Kingdom, and the United States. Emerging market economies consist of the following 22 economies: Argentina, Brazil, Chile, China, Colombia, Egypt, Hong Kong SAR, Indonesia, India, Israel, Korea, Morocco, Mexico, Malaysia, Pakistan, Peru, Philippines, Singapore, Thailand, Turkey,

Venezuela, and South Africa. Aggregate capital flows for any set of countries are calculated by summing up the values for individual countries in the set.)

The first part of this article, which draws from joint work (Ju and Wei 2006), provides an analytical perspective on the volume of international capital flows, which can be regarded as either too low (known as the Lucas paradox) or too high (when compared with the logic of factor price equalization). The second part, which draws from a different set of recent work (Prasad et al. 2003; Wei 2006; Kose et al. 2006), examines some apparent mismatch between theory and empirics on the economic consequences of international capital flows, and discusses ways to reconcile them.

The Volume of International Capital Flows: Paradoxes and Possible Solutions

The extent of cross-border capital movement can be measured by flows at a given point in time or by stocks accumulated over time. Capital inflows are net purchases of domestic assets by foreign residents, whereas capital outflows are net purchases of foreign assets by domestic residents. These data are well described in the International Monetary Fund's Balance of Payments statistics. For stock data, the IMF reports information for a few countries in recent years. Lane and Milesi-Ferreti (2001, 2005) expand the country and year coverage by combining this information with cumulative flows adjusted for valuation effects.

A country's exposure to international capital flows can be measured either by its government's policies (restrictions or incentives vis-à-vis capital flows) or by the actual amount of capital movement (scaled by the size of the recipient economy). The latter, the *de facto* measure, does not need to agree with the former, the *de jure* measure. For example, some countries may have many legal restrictions on capital movement (and hence a low exposure to capital flows by the *de jure* measure), but massive capital flight (and hence a high exposure by the *de facto* measure). A practical *de facto* measure of a country's exposure to cross-border capital movement is the sum of the country's total foreign assets and total foreign liabilities, divided by the country's GDP. For some economic questions, such as the effect of international capital flows on economic growth, the *de facto* measure may be more meaningful than the *de jure* measure.

Is the volume of capital flows observed in the data consistent with economic theory? Using a one-sector model, Lucas (1990) argues that it is a paradox that more capital does not flow from rich to poor countries. His reasoning goes as follows. Let $y = f(L, K)$ be a constant-returns-to-scale production function, where y is the output produced using labour L and capital K. Let p be the price of the good, and w and r be the returns to labour and capital, respectively. Firm's profit maximization problem gives

$$r = p\partial f(L,K)/\partial K = p\partial f(1,K/L)/\partial K \qquad (1)$$

If the product price is equalized across countries under free trade, the law of diminishing marginal product implies that r is higher in the country with a lower capital–labour ratio. As an illustration, Lucas calculates that the return to capital in India should be 58 times as high as that in the United States based on their factor endowment. Facing a return differential of this magnitude, one should observe a lot more capital flowing from rich to poor countries. That too little is observed in the data has come to be known as the 'Lucas paradox'.

Lucas (1990) discusses three possible explanations (within a one-sector framework): (*a*) a worker in a rich country could be several times more productive than her counterpart in a poor country; (*b*) human capital may be a missing factor and is likely much higher in a rich country; and (*c*) political risk and hence the required risk premium may be substantially higher in a poor country. Reinhart and Rogoff (2004) illustrate the last point for a set of countries with frequent default on their external debt.

Lucas's logic can be turned on its head in a multi-sector model. More precisely, in a standard Heckscher–Ohlin–Samuelson model with two

goods, two factors, and two countries, firms earn zero profit. So one must have:

$$p_1 = c_1(w, r) \text{ and } p_2 = c_2(w, r) \qquad (2)$$

where $c(.)$ is the unit cost function and the numerical subscripts represent sectors. This implies that the factor prices are uniquely determined by product prices, and are independent of factor endowments. Since free trade in goods equalizes the product prices across countries, factor returns must also be equalized even in the absence of cross-border capital and labour movement. This was first pointed out by Samuelson (1948) and has become known as the 'factor price equalization theorem'. Two countries with different capital–labour ratios would simply produce different mixes of outputs, but the marginal returns to physical capital are the same everywhere. In other words, zero capital flow is needed in equilibrium. This is true with or without cross-country differences in effective labour, human capital or political risk. The actual capital flow appears excessive on this logic.

One might think that the theorem of factor price equalization is too naive, requiring restrictive assumptions that surely do not hold in a more realistic setting with many countries, goods and factors. However, Ju and Wei (2006) show that, in a generalized neoclassical framework, relatively weak conditions are sufficient for factor prices to be equalized across countries (without factor movement). In particular, while the United States and India may not appear to satisfy the conditions for the factor prices to be equalized between them in a two-country model, it is nonetheless possible for factor prices to be equalized through a chain of country pairs (for example, the United States and Spain, Spain and Greece, Greece and Thailand, and Thailand and India). This means that it may be more difficult than it first appears to escape from the logic of factor price equalization within a neoclassical framework, and that free trade in goods can completely substitute for capital mobility.

Obstfeld and Rogoff (2001) proposed that the existence of trade costs could explain the low but positive international capital flow. Trade costs do break factor price equalization even in a two-sector, two-factor model. However, as tariffs and transport costs decline over time, factor prices (including returns to capital) should converge across countries. This should lead to a decline in international capital flow (by the logic of factor price equalization), which is contradicted by the data.

Cross-country differences in total factor productivity (TFP) is another influential explanation of the Lucas paradox. While the discussion is usually couched in a one- sector model, it *could* work even in a multi-sector model. In particular, in a two-sector model, if the TFPs in both sectors are many times higher in the United States than in India, then return to capital in the United States could be only slightly lower than in India, justifying the observed small amount of capital flow. What drives the TFP differential across countries can be the quality of institutions, including the protection of property rights and the control of bureaucratic corruption. However, the TFP story can also go in the opposite direction in principle, exacerbating rather than resolving the Lucas paradox. In particular, if the United States has a greater TFP advantage in the labour-intensive sector than in the other sector, then this could further depress the return to capital from what already results from a high capital–labour ratio. This suggests that one has to be precise about the nature of the TFP differences in order to deliver predictions on the sign and the size of international capital flows.

Moving outside the neoclassical box, Ju and Wei (2006) introduce financial contracts and heterogeneous firms into an otherwise standard two-sector, two-factor framework. A key implication of the model is the separation between return to physical capital and return to financial investment. In particular, India could have a high return to physical capital due to its relatively low capital–labour ratio, but a low return to financial investment due to its relatively inefficient financial system. In addition, heterogeneous firms give rise to diminishing marginal returns at the sector level even though every firm has a constant returns technology. As a result, factor price equalization (before factor movement) does not hold in

this model. In equilibrium, it is possible for financial capital to leave India for the United States, and for physical investment to flow in the reverse direction, resulting in a moderate amount of net flow. In this model, the return to capital (before capital flows) is still higher in India (with a lower capital-to-labour ratio) than the United States, but the differential in return is much smaller than in a one sector model. Thus, Ju and Wei's non-neoclassical two-sector, two-factor model partially restores the result of a typical one-sector model (that is, return to capital is determined in part by factor endowment) but does not generate the Lucas paradox.

Effects of International Capital Flows on Economic Growth

The Gap Between Theories and Empirics

International capital flows have the potential to bring a variety of benefits to recipient countries. In theory, financial globalization could raise a country's economic growth rate through a number of direct and indirect channels.

The direct channels include (*a*) augmenting domestic savings, (*b*) reducing the cost of capital through better allocation of risks (Henry 2000; Stulz 1999), (*c*) transferring technology and managerial know-how (Grossman and Helpman 1991), and (*d*) stimulating development of the domestic financial sector (Levine 1996; 2005). The indirect channels include (a) promoting specialization (Brainard and Cooper 1968; Imbs and Wacziang 2003), and (b) committing to better economic policies (Gourinchas and Jeanne 2004; Tytell and Wei 2004).

Yet a massive body of empirical papers has often found mixed results, suggesting that the benefits are not straightforward. Kose et al. (2006) survey 20 scholarly articles written between 1994 and 2005 that have empirically estimated the effect of exposure to international capital flows on economic growth. A majority of these papers (16 out of 20) find no, or at best mixed, effects. This echoes the conclusion in earlier survey articles by Eichengreen 2001 and Prasad et al. (2003) that it is not easy to find a strong and robust causal effect from financial globalization to economic growth, especially for developing countries.

Indeed, one alleged source of collateral damage of financial globalization is an increased propensity for developing countries to experience currency crises or other types of financial turmoil. For example, while the pace of cross-border capital flows picked up in the 1980s, there have also been more financial crises since around 1990, including the crises in Mexico in 1994, the Asian financial crisis during 1997–9, the Russian meltdown in 1999, and the Argentinean and Uruguayan crises of 2001–2. Most such crises tend to set countries back in their growth aspirations for a number of years.

Reconciling Theories with Empirical Patterns

Financial crises do not prove that financial integration is a bad thing. Indeed, almost all developed countries are financially integrated, and very few developing countries, once embarked on a path of integration, would go back to financial isolation. So why do countries aspire to become financial integrated and yet experience so many bumps and potholes along the way? The literature has proposed *independently* two views: a composition hypothesis and a threshold hypothesis.

The composition hypothesis maintains that not all capital flows are equal. International direct investment, and perhaps international portfolio flows, appear to be robustly associated with a positive effect on economic growth (Borensztein et al. 1998; Bekaert et al.2005). In contrast, there is no strong evidence that private foreign debt including international lending has robustly promoted economic growth. Indeed, one sometimes finds evidence that international lending is negatively associated with economic growth. Official aid flows do not robustly support growth either (Rajan and Subramanian 2005).

Composition of capital flows has also been related to a country's propensity to experience a currency crisis. In their study of all episodes of currency crises in emerging markets during 1971–92, Frankel and Rose (1996) report that, while virtually no variable has a strong predictive power for subsequent currency crashes, the

composition of capital inflows is one of the very few variables that are robustly related to the probability of a currency crisis. In particular, the share of foreign direct investment (FDI) in a country's total capital inflow is negatively associated with the probability of a currency crisis. This is confirmed in several subsequent studies including Frankel and Wei (2005). Other dimensions of composition are the maturity structure of external debt (the greater the share of short-term debt, the more likely a crisis), and the currency denomination of external debt (the greater the share of foreign currency debt, the more likely a crisis) (Frankel and Rose 1996; Radelet and Sachs 1998; Rodrik and Velasco 1999).

The threshold hypothesis states that certain minimum conditions have to be met before a country can be expected to benefit from financial globalization. Otherwise, the country could experience more crises and lower growth. The threshold effect comes in various versions. Only countries with reasonably good public institutions (for example, adequate control of corruption) and a minimum level of human capital seem to be able to translate exposure to financial globalization into stimulus to investment and growth on a sustained basis (see the surveys by Prasad et al. 2003; Kose et al. 2006). It is not difficult to imagine why countries with weak institutions may not benefit from financial globalization. In a highly corrupt country, for example, more capital inflows are likely to result in more consumption by a few elite families or in bigger Swiss bank accounts rather than more productive investment. So more capital flows may not result in higher growth rates. If capital inflows help to promote excessively risky projects backed by governments, then more inward capital flows could translate into an increased probability of a financial crisis.

Is the Composition Effect a Consequence of the Threshold Effect?

Rather than viewing the threshold effect and the composition effect as two *rival* hypotheses, Wei (2000a, b, 2001) suggests a concrete connection between the two: countries with better public institutions are likely to attract more international direct investment than international bank loans. Wei derives evidence from data on bilateral FDI reported by OECD source countries, and bilateral international lending reported by Bank for International Settlements (BIS) member countries. In the earlier work, Wei measures quality of public institutions by perception of corruption reported in surveys of firms such as those conducted by the World Economic Forum for its *Global Competitiveness Report* or by the World Bank for its *World Development Report*.

Recent evidence on investment by international mutual funds suggests that better institutions measured by a high degree of government and corporate transparency help to attract more international equity investment than that predicted by the international capital asset pricing model (ICAPM) (Gelos and Wei 2005). So the composition effect and the threshold effect are perhaps just the two sides of the same coin.

Not everyone has found the same result. Hausmann and Fernandez-Arias (2000) report no relationship between share of FDI in total capital inflows and good institutions. In a panel of advanced and developing countries, Albuquerque (2003) finds the share of FDI in total inflows to be negatively related to good credit rating. It is important to note that Albuguerque's measure is about financial development rather than quality of public institutions generally, whereas Haumann and Fermandez-Arias mix measures of financial development and property rights institutions. As Ju and Wei (2006) point out, financial development and quality of public institutions have different effects, in theory, on the composition of capital flows. Furthermore, none of these studies employs instrumental variables to correct for possible measurement errors and endogeneity of the corruption or other institutional measures.

In any case, more recent papers with an instrumental variable approach and arguably better data again affirm the earlier conclusion that there may be an intimate relationship between the institutional threshold effect and the composition effect. Using data from the IMF on balance of payments, Alfaro et al. (2005) find that good institutional quality is a key determinant of total capital inflows. Papaioannou (2005) reports that foreign asset holdings by BIS banks, including their

portfolio assets and direct investments, tend to be higher in destinations with better institutions.

Using recently available data from the IMF on member countries' international investment position (IMF 2002), Faria and Mauro (2004) present evidence that countries with strong institutions are likely to attract more equity-like capital flows (FDI and portfolio equity flows) than other types of capital. Their measure of Einstitutional quality is the average of six indicators – voice and accountability, political stability and absence of violence, government effectiveness, regulatory quality, rule of law, and control of corruption – as computed and reported by Kaufmann et al. (2003). An important feature of the study is that the authors address explicitly the possibility that the composite institutional index may be measured with errors and/or be endogenous. They employ as instrumental variables log settler mortality during the early colonial period as proposed by Acemoglu et al. (2001) and ethno-linguistic fragmentation first used by Mauro (1995). The instrumental variable approach reaffirms their basic conclusion.

Wei (2006) furnishes evidence that the effects of the quality of public institutions and the level of financial development can indeed be different. In particular, weak public institutions strongly discourage FDI, and possibly foreign debt, as shares of a country's total foreign liabilities, but appear to encourage borrowing from foreign banks. In comparison, low financial-sector development discourages inward portfolio equity flows but encourages inward FDI. The finding that poor financial development could encourage FDI may sound surprising. A possible story is set out in Ju and Wei (2006). Essentially, in countries with poor financial systems but also low capital–labour ratios, the return to financial capital is low. Hence domestic households would want to take savings out of the country, and international portfolio investors do not wish to come in. As the same time, as long as the risk of expropriation is not too high, the depression of domestic investment due to poor domestic financial development could raise the return to FDI.

To gain confidence that these patterns reflect causal relations, Wei (2006) employs instrumental variables for the institutional measures based on the economic histories of the countries in the sample, in particular the log mortality rate of the European settlers in former colonies à la Acemoglu et al. (2001), and the origin of legal systems à la La Porta et al. (1998). The instrumental variable approach bolsters the argument that weak institutions are a *cause* of the unfavourable composition of capital inflows.

To summarize, the cumulative evidence points to the strong possibility that weak public institutions tilt the composition of capital flows into a country away from FDI and portfolio equity flows and towards debt, including bank loans, making the country more vulnerable to a currency crisis and less able to translate a given amount of capital inflow into stimulus for economic growth. While the composition and the threshold effects are not identical, they are very likely related.

For an institutionally challenged country, more research is needed to determine whether it should wait for its institutions to be sufficiently improved before opening up to global capital flows, or use exposure to the international capital market as a disciplinary device to improve its institutions.

See Also

- ▶ Financial Liberalization
- ▶ Foreign Direct Investment
- ▶ Globalization

Acknowledgment *The author is grateful to Andre Faria, Jiandong Ju, Ayhan Kose, Paolo Mauro, Gian Maria Milesi-Ferretti, Romain Ranciere, Kenneth Rogoff, and Irina Tytell for helpful discussion, and Yuanyuan Chen and Patricia Medina for capable research and editorial assistance. The views expressed in this article are the author's own, and do not reflect those of the IMF or any other organization he is associated with.*

Bibliography

Acemoglu, D., S. Johnson, and J. Robinson. 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review* 91: 1369–1401.

Albuquerque, R. 2003. The composition of international capital flows: Risk sharing through foreign direct

investment. *Journal of International Economics* 1: 353–383.
Alfaro, L., S. Kalemli-Ozcan, and V. Volosovych. 2005. Why doesn't capital flow from rich to poor countries? An empirical investigation. Working paper no. 11901. Cambridge, MA: NBER.
Bekaert, G., C. Harvey, and C. Lundblad. 2005. Does financial liberalization spur growth? *Journal of Financial Economics* 77: 3–56.
Borensztein, E., J. de Gregorio, and J.-W. Lee. 1998. How does foreign direct investment affect economic growth? *Journal of International Economics* 45: 115–135.
Brainard, W., and R. Cooper. 1968. Uncertainty and diversification of international trade. *Food Research Institute Studies in Agricultural Economics, Trade, and Development* 8: 257–285.
Caballero, R., E. Farhi, and P.-O. Gourinchas. 2005. An equilibrium model of global imbalances and low interest rate. Working paper. MIT and UC Berkeley.
Eichengreen, B. 2001. Capital account liberalization: What do cross-country studies tell us? *World Bank Economic Review* 15: 341–365.
Faria, A., and P. Mauro. 2004. Institutions and the external capital structure of countries. Working paper no. 04/236. Washington, DC: IMF.
Frankel, J., and A. Rose. 1996. Currency crashes in emerging markets: An empirical treatment. *Journal of International Economics* 41: 351–366.
Frankel, J., and S.-J. Wei. 2005. Managing macroeconomic crises: Policy lessons. In *Managing economic volatility and crises: A practitioner's guide*, ed. J. Aizenman and B. Pinto. New York: Cambridge University Press.
Gelos, G., and S.-J. Wei. 2005. Transparency and positions of international mutual funds. *Journal of Finance* 60: 2987–3020.
Gourinchas, P.-O., and O. Jeanne. 2004. On the benefits of capital account liberalization for emerging economies. Working paper. University of California, Berkeley, and the International Monetary Fund.
Grossman, G., and E. Helpman. 1991. Trade, knowledge spillovers, and growth. *European Economic Review* 35: 517–526.
Hausmann, R., and E. Fernandez-Arias 2000. Foreign direct investment: Good cholesterol? Working paper no. 417. Research Department, Inter-America Development Bank.
Henry, P. 2000. Stock market liberalization, economic reform, and emerging market equity prices. *Journal of Finance* 55: 529–564.
Imbs, J., and R. Wacziang. 2003. Stages of diversification. *American Economic Review* 93: 63–86.
IMF (International Monetary Fund). 2002. International investment position: A guide to data sources. Online. Available at www.imf.org/external/np/sta/iip/guide/index.htm. Accessed 11 May 2006.
Ju, J., and S.-J. Wei. 2006. A solution to two paradoxes of international capital flows. NBER working paper, forthcoming. Cambridge: National Bureau of Economic Research.
Kaufmann, D., A. Kraay, and M. Mastruzzi. 2003. Governance matters III: Governance indicators for 1996–2002. Policy research working paper no. 3106. Washington DC: The World Bank.
Kose, M., E. Prasad, K. Rogoff, and S.-J. Wei. 2006. Financial globalization: A reappraisal and synthesis. Prepared for the *Journal of Economic Literature*.
La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny. 1998. Law and finance. *Journal of Political Economy* 106: 1113–1155.
Lane, P., and G. Milesi-Ferreti. 2001. The external wealth of nations: Measures of foreign assets and liabilities for industrial and developing nations. *Journal of International Economics* 55: 263–294.
Lane, P., and G. Milesi-Ferreti. 2005. The external wealth of nations mark II: Revised and extended estimates of foreign assets and liabilities, 1970–2004. Working paper no. 06/69. Washington, DC: IMF.
Levine, R. 1996. Foreign banks, financial development, and economic growth. In *International financial markets: Harmonization versus competition*, ed. C. Barfield. Washington, DC: AEI Press.
Levine, R. 2005. Finance and growth: Theory and evidence. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.
Lucas, R. 1990. Why doesn't capital flow from rich to poor countries? *American Economic Review* 80: 92–96.
Mauro, P. 1995. Corruption and growth. *Quarterly Journal of Economics* 110: 681–712.
Obstfeld, M., and K. Rogoff. 2001. The six major puzzles in international macroeconomics: Is there a common cause? *NBER Macroeconomics Annual* 15: 339–390.
Papaioannou, E. 2005. What drives international bank flows? Politics, institutions and other determinants. Working paper no. 437. Frankfurt: European Central Bank.
Prasad, E., K. Rogoff, S.-J. Wei, and M. Kose. 2003. Effects of financial globalization on developing countries: Some empirical evidence, Occasional paper no. 220. Washington, DC: IMF.
Radelet, S., and J. Sachs. 1998. The east Asian financial crisis: Diagnosis, remedies, and prospects. *Brookings Papers on Economic Activities* 1998 (1): 1–74.
Rajan, R., and A. Subramanian. 2005. What undermines aid's impact on growth? Working paper no. 05/126. Washington, DC: IMF.
Reinhart, C., and K. Rogoff. 2004. Serial default and the 'paradox' of rich to poor capital flows. *American Economic Review* 94 (2): 52–58.
Rodrik, D., and A. Velasco. 1999. Short-term capital flows. Working paper no. 7364. Cambridge, MA: NBER.
Samuelson, P. 1948. International trade and equalization of factor prices. *Economic Journal* 58: 163–184.
Smarzynska, B., and S.-J. Wei 2000. Corruption and the composition of foreign direct investment: Firm-level evidence. Working paper no. 7969. Cambridge, MA: NBER.

I

Stulz, R. 1999. Globalization of equity markets and the cost of capital. Working paper no. 7021. Cambridge, MA: NBER.
Tytell, I., and S.-J. Wei. 2004. Does financial globalization induce better macroeconomic policies? Working paper no. 04/84. Washington, DC: IMF.
Wei, S.-J. 2000a. How taxing is corruption on international investors? *Review of Economics and Statistics* 82: 1–11.
Wei, S.-J. 2000b. Local corruption and global capital flows. *Brookings Papers on Economic Activities* 2000 (2): 303–354.
Wei, S.-J. 2001. Domestic crony capitalism and international fickle capital: Is there a connection? *International Finance* 4: 15–45.
Wei, S.-J. 2006. Connecting two views on financial globalization: Can we make further progress? Working paper.

International Coordination in Asylum Provision

Yuji Tamura

Abstract

This article summarises theoretical studies on asylum provision in multi-country settings. The common feature of their models is the assumption that asylum-related policies of safe countries generate cross-border externalities. The presence of externalities results in inefficiently low provision of asylum. The studies explore ways to increase asylum provision to the efficient level, but reveal more difficulties than a solution.

Keywords

Asylum provision; Asylum seeker; Cross-border externalities; International coordination; International migration; International public goods; Refugee; Refugee protection

JEL Classifications

F22; F53; H87; O15

This article summarises the current state of the theoretical literature on the provision of asylum – the protection of refugees by sheltering them within the provider's territories. The existing studies offer insight into the incentive problem of asylum provision when there is more than one potential host country. The studies show that asylum provision remains inefficiently low even if people in safe countries are humanitarian and care about the welfare of refugees. This is because of the way people in safe countries benefit from asylum provision in the models. The government of a safe country derives a benefit from the protection of refugees because it cares about the welfare of its own citizens who in turn care about the welfare of refugees abroad. However, this benefit accrues to the citizens whether the protection is provided in their own country or in other safe countries. Since hosting refugees is costly, each safe country has an incentive to rely on the asylum provision of other safe countries. Consequently, the provision of asylum remains inefficiently low.

This type of incentive problem in asylum provision was first articulated verbally before the model-based studies appeared, e.g. Suhrke (1998). As a matter of fact, the free-or easy-riding problem (Cornes and Sandler 1996) is not specific to asylum provision but is common to the provision of various public goods that are distinguished from private goods by the degree of non-rivalry and non-excludability in consumption. In our context, the 'consumption' of refugee protection is not rivalrous because when the humanitarian citizens of one country gain from the protection of refugees their gain does not reduce the benefit that accrues to the humanitarian citizens of another country. The 'consumption' of asylum is not excludable either. The provider cannot prevent humanitarian citizens of other countries from enjoying the protection of refugees. The provision of a public good thus generates positive externalities – the benefits enjoyed by people who did not supply the good.

Using an overarching framework where asylum enters as an international public good, we look at how different factors influence asylum provision. The existing studies examine how we could mitigate the incentive problem among potential host countries. However, their searches for ways to increase asylum provision to the

efficient level reveal, instead of a solution, difficulties in resolving the incentive problem. This article omits discussion of the literature in relation to policy coordination in reality, as the non-technical companion article (Suriyakumaran and Tamura 2016) includes it.

Benchmark Model

Consider a fixed number of safe countries indexed by $n = 1, \ldots, N$. They face a mass of identical refugees. Asylum is modelled as an international public good.

Each country's net benefit is defined as follows:

$$u_n(a_n, a_{-n}) = b_n(a_n, a_{-n}) - c_n(a_n)$$

where $a_n \geq 0$ denotes the number of refugees that country n decides to host and $a_{-n} \geq 0$ denotes the number of refugees that the $N-1$ other countries decide to host. The benefit function, $b_n(\cdot)$, is increasing in asylum provision by any country. Here, a_n and a_{-n} are entered separately to allow for the possibility that asylum is an impure public good. That is, a safe country might well regard a foreign country's provision of asylum as an imperfect substitute of asylum provision by itself, although one foreign country's provision is assumed to be a perfect substitute of another foreign country's provision. Reasons for considering asylum provision as an impure public good include the importance of doing it yourself as opposed to reliance on other safe countries (self-satisfaction aspect) and the generation of international prestige through the exhibition of altruistic actions (Andreoni 1989). The cost function, $c_n(\cdot)$, is increasing only in its provision of asylum, as the cost of providing a_{-n} is borne by the other safe countries. The net benefit, $u_n(\cdot)$, is concave in a_n, i.e. $\partial^2 b_n / \partial \alpha_n^2 < 0$ and $c_n'' \geq 0$, or $\partial^2 b_n / \partial \alpha_n^2 < 0$ and $c_n'' \geq 0$.

The absence of international coordination is modelled as a static game of complete information in which each country independently maximises its net benefit by taking others' asylum provision as given and choosing its provision of asylum, i.e.

$$\max_{a_n} u_n(a_n; a_{-n}).$$

The resulting Nash equilibrium is compared with the outcome of

$$\max_{\alpha_1, \ldots, \alpha_N} \sum_{n=1}^{N} u_n(a_n, a_{-n}).$$

Under international coordination, the countries are thus assumed to maximise the utilitarian welfare together. This internalises the positive externalities generated by each country's provision of asylum.

Hatton (2004) analyses this model for two identical safe countries, i.e. $b_1(a_1, a_2) = b_2(a_2, a_1)$ and $c_1(a_1) = c_2(a_2)$. He shows that the internalisation of the positive externalities increases the total provision of asylum by increasing provision in both countries. Since each country's benefit function increases in the total provision while the cost function increases only in its own provision, the net benefit is greater under the utilitarian maximisation than in the Nash equilibrium. Hence the identical host countries should be willing to coordinate to maximise their total welfare, resulting in a Pareto efficient outcome.

This conclusion does not require two potential host countries to be identical. Czaika (2009) shows that both countries can benefit from the utilitarian maximisation as long as they are sufficiently similar to each other in their benefit and cost parameters. In other words, when the countries are sufficiently different from each other, the utilitarian maximisation requires one of the two to be worse off than in the Nash equilibrium. The total welfare maximisation is not beneficial for the country where asylum provision is very costly and/or is not sufficiently appreciated.

Role of Cross-Border Financial Transfers in Increasing Asylum Provision

Facchini et al. (2006) and Czaika (2009) examine the role of financial transfers in enabling two

non-identical countries to increase the total provision of asylum. While Facchini et al. (2006) simply assume that two countries agree to share the total net benefit equally through a financial transfer from one country to another, Czaika (2009) allows two countries to decide on their cross-border financial transfers in a two-stage game with complete and imperfect information. The countries first decide on their transfers simultaneously, and then decide on asylum provision simultaneously. He shows that the total provision of asylum in the subgame-perfect Nash equilibrium can be greater than that in the Nash equilibrium without a financial transfer. The country that benefits more from asylum provision abroad than from domestic provision (in other words, the marginal cost of asylum provision is high) is willing to financially support the other country which benefits more from domestic provision than from provision abroad. However, the possibility of financial transfers is shown to be insufficient to achieve the total welfare maximisation.

Allocating Tradable Asylum Quotas to Achieve the Utilitarian Welfare Maximisation

Fernández-Huertas Moraga and Rapoport (2014, 2015) show that a fixed number of potential host countries can maximise the total net benefit if they can agree with an initial allocation of tradable asylum quotas that can total the utilitarian welfare-maximising level of provision. In this competitive scheme, originally proposed verbally by Schuck (1997), the equalisation of marginal costs of asylum provision across the participating countries determines the price of transferring to another country the obligation to grant asylum to a refugee. Countries that can provide asylum only at high costs reduce their quotas by paying other low-cost countries to host refugees in excess of their initial quotas.

For the countries to willingly participate in the scheme, the initial allocation of tradable quotas must be incentive-compatible. Other things being equal, the scheme should initially allocate larger quotas (larger than the post-trading quotas that achieve the utilitarian maximisation) to countries that derive greater externality-related benefits from the scheme (that is, the externality-related benefit is greater under the scheme than in the Nash equilibrium without tradable quotas). Also, other things being equal, smaller quotas (smaller than the post-trading quotas that achieve the utilitarian maximisation) should be initially allocated to countries that bear higher costs under the scheme (that is, the cost of asylum provision is higher under the scheme than in the Nash equilibrium without tradable quotas). An initial allocation that meets these incentive compatibility constraints enables all participating countries to increase their net benefits through quota trading.

Fernández-Huertas Moraga and Rapoport (2014, 2015) point out the difficulty of identifying incentive compatibility constraints for deriving a correct initial allocation. Relevant benefit and cost parameters are likely to be private information, and a potential participant may lack an incentive to reveal true parameters because the initial allocation can be manipulated to its advantage by giving false parameters.

Fernández-Huertas Moraga and Rapoport (2014, 2015) also show that individual preferences of refugees over the potential host countries and the preferences of those countries over refugees with different characteristics can be taken into consideration through matching mechanisms. For example, in a deferred acceptance algorithm (Roth 1985), safe countries offer asylum to their most preferred refugees, and each of the refugees either accepts one of the received offers or declines all offers according to her/his preference. The process repeats with the remaining refugees until all tradable quotas are met. As we focus on the welfare of host countries in this literature, this algorithm may be attractive because safe countries are likely to benefit from an efficiency gain by filling their asylum quotas with preferred refugees. To justify the use of a matching mechanism, the efficiency gain must outweigh a potential increase in uncertainty that results from untruthful preference revelation by participating countries (Roth 1985).

Introducing Asylum-Seeking Behaviour

So far we have assumed a mass of identical refugees facing a fixed number of safe countries, and each country decides on how many of them to host. Refugees do not have a choice. The assumption seems reasonable for analysing resettlement situations where the mass consists of refugees who are similar to each other in hosting costs-related characteristics (such as language skill and ethnicity) and in the evidence of protection need which every safe country can easily verify.

We now introduce asylum-seeking behaviour. Each refugee chooses whether and where to seek asylum. We adapt Monheim-Helstroffer and Obidzinski's (2010) approach to our framework. (Their model does not regard asylum as an international public good.) Consider two safe countries facing a mass of refugees normalised to one. We assume bi-dimensional heterogeneity among the refugees. They differ in the strength of evidence of protection need and are uniformly distributed over the interval [0, 1]. They also differ in the preferred destination. A fraction, h, of the unit mass prefer to seek asylum in country 1 if they have evidence to exceed the standards of proof in both countries. The rest prefer country 2. The preference parameter, h, summarises the heterogeneity across refugees in all dimensions except the ability to meet each country's standard of proof. For example, a refugee may prefer country 1 to country 2 partly because seeking asylum in country 1 is financially less costly than doing so in country 2. A refugee's preference may also be influenced by the cross-country differences in the existing social network. Let us assume $h \in \left(0, \frac{1}{2}\right)$, i.e. country 2 is more popular than country 1 among the refugees. Each country sets a standard of proof, $s_n \in [0; 1]$, and accepts asylum seekers whose evidence exceeds the standard. Refugees decide whether and where to seek asylum in response to the standards of proof. For simplicity, refugees are assumed unconstrained in all dimensions (such as finance and time) other than the ability to satisfy each country's standard of proof. (See also Giordani and Ruta (2013) who similarly model the destination choice of migrants in response to immigration policies in a multi-country setting, although their study does not deal with refugees.)

To understand asylum seeking behaviour, suppose $s_1 < s_2$, i.e. the more popular country sets a higher standard of proof. Then the unit mass consists of the following four groups:

- s_1 refugees who do not seek asylum because their evidence is weak
- $s_2 - s_1$ refugees who seek asylum in country 1, including $(s_2 - s_1)(1 - h)$ refugees who prefer country 2 but their evidence is not sufficiently strong
- $(1 - s_2)h$ refugees who seek asylum in country 1
- $(1 - s_2)(1 - h)$ refugees who seek asylum in country 2

By applying the same reasoning to the case of $s_1 \geq s_2$, we obtain each safe country's asylum provision as a function of the standards of proof as follows:

$$a_1(s_1, s_2) = \begin{cases} h - s_1 + (1 - h)s_2 & \text{if } s_1 < s_2 \\ (1 - s_1)h & \text{otherwise} \end{cases}$$

$$a_2(s_1, s_2) = \begin{cases} (1 - h) - s_2 + hs_1 & \text{if } s_1 < s_2 \\ (1 - s_1)(1 - h) & \text{otherwise} \end{cases}$$

Each country takes these equations into account in maximising its net benefit with respect to the standard of proof. The equations imply that a country can reduce its asylum provision by increasing its standard of proof. They also imply that a country has to increase its asylum provision when the other country's standard of proof is higher than its standard and increases. Thus, setting a standard higher than the other country's standard is equivalent to redirecting some refugees to the other country. By explicitly incorporating asylum-seeking behaviour, we thus introduce cross-border externalities that are distinct from the positive externalities that result from the assumption that asylum is an international public good.

By substitution, each country's net benefit is now written as

$$u_n(a_n, a_{-n}) = \begin{cases} b_n(a_n(s_n, s_{-n}), a_{-n}(s_{-n})) - c_n(a_n(s_n, s_{-n})) & \text{if } s_n < s_{-n} \\ b_n(a_n(s_n), a_{-n}(s_n, s_{-n})) - c_n(a_n(s_n)) & \text{otherwise} \end{cases}$$

for $n, -n = 1; 2$ and $n \neq -n$: The Nash equilibrium, (s_1^*, s_2^*), is characterised by

$$\frac{\partial b_n}{\partial a_n}\left(a_n\left(s_n^*, s_{-n}^*\right), a_{-n}\left(s_{-n}^*\right)\right) = c_n'\left(a_n\left(s_n^*, s_{-n}^*\right)\right),$$

$$\frac{\partial b_{-n}}{\partial a_{-n}}\left(a_{-n}\left(s_{-n}^*\right), a_n\left(s_n^*, s_{-n}^*\right)\right) = c_n'\left(a_{-n}\left(s_n^*\right)\right) + \frac{\partial b_{-n}}{\partial a_{-n}}\left(a_{-n}\left(s_{-n}^*\right), a_n\left(s_n^*, s_{-n}^*\right)\right)$$

$$s_n^* < s_{-n}^*,$$

for $n, -n = 1; 2$ and $n \neq -n$; or

$$\frac{\partial b_1}{\partial a_1}\left(a_1\left(s_1^*\right), a_2\left(s_1^*, s_2^*\right)\right) = c_1'\left(a_1\left(s_1^*\right)\right) + \frac{\partial b_1}{\partial a_2}\left(a_1\left(s_1^*\right), a_2\left(s_1^*, s_2^*\right)\right),$$

$$\frac{\partial b_2}{\partial a_2}\left(a_2\left(s_2^*\right), a_1\left(s_1^*, s_2^*\right)\right) = c_2'\left(a_2\left(s_2^*\right)\right) + \frac{\partial b_2}{\partial a_1}\left(a_2\left(s_2^*\right), a_1\left(s_1^*, s_2^*\right)\right),$$

$$s_1^* = s_2^*.$$

Note that, say, for $s_1 < s_2$, country 1's first-order condition is not different from the case without asylum-seeking behaviour. That is, the marginal benefit and marginal cost of providing asylum in the country are equalised. This is because country 1 cannot influence country 2's asylum provision so long as s_1 remains lower than s_2.

However, country 2's first-order condition contains an extra term on the right-hand side. It is the marginal benefit that country 2 gains from country 1's asylum provision. Country 1's asylum provision increases when country 2 raises its standard of proof because the refugees who become disqualified in country 2 at the margin will successfully seek asylum in country 1. This additional term on the right-hand side of the first-order condition motivates country 2 to set a high standard that results in the country's asylum provision being lower than without asylum-seeking behaviour.

We can easily confirm that the total provision of asylum in the Nash equilibrium is lower than that under utilitarian maximisation. In other words, total welfare maximisation requires lower standards of proof than in the Nash equilibrium. Whether $s_1 < s_2$ or not, the utilitarian maximisation is characterised by

$$\frac{\partial b_1}{\partial a_1}(a_1(\cdot), a_2(\cdot)) + \frac{\partial b_2}{\partial a_1}(a_2(\cdot), a_1(\cdot)) = c_1'(a_1(\cdot)),$$

$$\frac{\partial b_2}{\partial a_2}(a_2(\cdot), a_1(\cdot)) + \frac{\partial b_1}{\partial a_2}(a_1(\cdot), a_2(\cdot)) = c_2'(a_2(\cdot)).$$

These two conditions show that both types of externality (cross-border benefits from asylum provisions and the cross-border deflection of asylum seekers through the requirement of high standards of proof) must be internalised to maximise total welfare. The internalised cross-border benefit is represented by the second term on the left-hand side of each equation. Recall that, in the Nash equilibrium with $s\ s_1^* < s_2^*$, for example, $\partial a_2/\partial a_1$ was a term on the right-hand side of the second condition, not a term on the left of the first condition, because country 2 used its standard to manipulate country 1's asylum provision to its own advantage.

The total welfare maximisation is incentive-compatible if each country's net benefit is higher than in the Nash equilibrium. However, in this extended framework, even if we assume that the two countries are homogeneous in terms of relevant benefit and cost parameters, i.e. $b(a_1, a_2) = b_2(a_2, a_1)$ and $c_1(a_1) = c_2(a_2)$, the destination preferences of refugees cause their total welfare-maximising standards of proof to diverge. Allowing more popular countries to set higher standards may be inadequate for non-economic reasons in dealing with refugees. Monheim-Helstroffer and Obidzinski (2010) examine two alternative coordination regimes: fixing an internationally common standard and setting the highest standard permitted internationally.

Under the common standard regime, all participating safe countries use the same standard of proof. That is, the standard of proof affects refugees' decisions on whether to seek asylum, but it does not affect the choice of destination among those who decide to seek asylum. The common standard regime removes deflection externalities, and asylum seekers choose their destinations solely according to their preferences. Monheim-Helstroffer and Obidzinski (2010) suggest that participation in the utilitarian maximisation by a common standard is always incentive-incompatible for more popular safe countries because they can increase their net benefits by deviating from the common standard upwards to deflect asylum seekers to less popular countries.

The other alternative coordination regime, the maximum standard regime, sets a standard that the participating safe countries are prohibited from exceeding. This approach leaves each country the room to choose its own standard, but the choice range is capped. The maximum standard is chosen by anticipating how the participating countries set their standards in response to the ceiling. That is, the maximum standard is the subgame-perfect Nash equilibrium of the two-stage game with complete and imperfect information. Like the common standard case, the total welfare-maximising ceiling prevents more popular countries from setting sufficiently high standards to achieve their net benefits as large as in the Nash equilibrium. However, the maximum standard regime is found superior to the common standard regime because the former allows less popular countries to maximise their net benefits with respect to their standards given more popular countries using the ceiling standard. Because the maximum standard regime forces more popular countries to provide more asylum than they would in the absence of the regime, less popular countries benefit through the increased provisions of asylum abroad. Less popular countries are able to use standards lower than the Nash equilibrium in the absence of the regime without attracting more asylum seekers to their countries.

When an Asylum Seeker Is Not Always a Refugee

The framework can easily accommodate the possibility that asylum seekers are not necessarily refugees – the concern held by many people in potential host countries. Suppose that, for any given evidence provided by an asylum seeker, the probability of the person being a true refugee is $r \in (0, 1)$. However, a safe country is unable to verify the falsity of evidence provided by non-refugees. Discounting evidence by probability is inappropriate, so each country hosts asylum seekers so long as they provide evidence (either genuine or falsified) that exceeds the required standard of proof. Let us assume that providing a non-refugee with asylum is as costly as providing a refugee with it, but generates no gross benefit. (Here we focus on non-refugee migrants who enter safe countries with false information, and we ignore the fact that economic migrants often benefit host countries.)

Assuming that the distribution of destination preferences of non-refugees is the same as that of refugees, each country's net benefit is now written as

$$u_n(g_n, a_n, g_{-n}, a_{-n}) = \begin{cases} b_n(g_n, (s_n, s_{-n}), g_{-n}, (s_{-n})) - c_n(a_n(s_n, s_{-n})) & \text{if } s_n > s_{-n} \\ b_n(g_n, (s_n), g_{-n}, (s_n, s_{-n})) - c_n(a_n(s_n)) & \text{otherwise} \end{cases}$$

for $n, -n = 1, 2$ and $n \neq -n$, where $g_n = ra_n$. Since every successful asylum seeker contributes to the benefit function by only a fraction and at the same time fully contributes to the cost function, the chosen standard of proof is higher than in the absence of non-refugee asylum seekers. This applies to both non-cooperative and utilitarian maximisation.

Bubb et al. (2011) were the first to introduce asylum-seeking by non-refugees in a multi-country setting. Their model is not as simple as the modified Monheim-Helstroffer and Obidzinski (2010) model given here. While I have simply assumed exogenous destination preferences, Bubb et al.'s non-refugees respond to international wage gaps and migration costs. The larger the wage gaps between potential destinations and refugee-generating countries and the lower the migration costs, the more economic migrants are tempted to seek asylum. To counter the incentive effects of increasing wage gaps and decreasing migration costs, safe countries increase their standards of proof. (Bubb et al. also assume that not all potential asylum seekers are fully aware of the strength of evidence they can produce, introducing uncertainty about the application outcome among applicants.)

Bubb et al. (2011) analyse a bilateral scheme by which wealthy safe countries counter the incentive effects and stop economic migrants from seeking asylum in them. In this scheme, originally proposed verbally by Hathaway and Neve (1997), safe-but-not-wealthy countries agree to provide asylum to people who have sought asylum in wealthy countries and have satisfied their standards of proof. The wealthy countries agree to pay the poor countries for hosting the successful applicants. In equilibrium, an economic migrant does not seek asylum because successful asylum-seeking results in migration from his/her country to another economically similar country. On the other hand, refugees continue to seek asylum, as the cost of remaining in their own countries is very high for them.

The scheme is likely to encourage refugees to seek asylum in poor safe countries because they know they will be sent to a poor country even if they meet the requirements in wealthy countries. By seeking asylum directly in poor safe countries, they are likely to minimise the costs involved in asylum-seeking, such as transport costs. An implication is that the scheme is incentive-incompatible for poor safe countries because wealthy countries financially compensate them for receiving people who have already been qualified in the wealthy countries, not for providing asylum to people who come directly to poor countries.

Conclusion

By applying the theory of private provisions of public goods to asylum provision by safe countries, the literature helps our understanding of the incentive problems involved. By incorporating asylum-seeking behaviour, we see how two different types of externality may work together to lower each country's provision: easy riding caused by positive externalities from public good provision by others and the use of strategically complementary policies to deflect asylum seekers to other safe countries. Studies of potential policy instruments – competitively tradable asylum quotas, harmonised standards of proof and bilateral contracts to pay to relocate qualified asylum seekers – reveal the weaknesses, as well as the strengths, associated with them as a solution to inefficiently low asylum provision. Hence they motivate us to examine how to deal with the weaknesses and also to come up with alternative schemes.

Researchers are likely to benefit from further extension of the existing framework and a search for a new framework. All existing studies pose asylum provision as a static game. (As we saw, Czaika (2009) and Monheim-Helstroffer and Obidzinski (2010) do present two-stage games,

but every country's asylum provision is simultaneously determined in one of the two stages.) However, in reality both applications for and provision of asylum take place over time. For example, a country's current provision of asylum may well influence the future applications that the country will receive. If so, forward-looking governments of safe countries must solve the game by anticipating future flows of humanitarian immigration. Foged and Peri (2016) present Danish evidence that municipalities with ethnic enclaves created by the refugee dispersal policy attracted further inflows of refugees in subsequent years, although not through asylum-seeking but via family reunification. It suggests that the destination preference of refugees depends at least partly on past asylum provisions.

As yet another refugee crisis has recently hit Europe, we realise that the problem of international coordination in asylum provision is a recurrent and unresolved one that demands further investigation.

See Also

- ▶ Economic Demography
- ▶ Exit and Voice
- ▶ External Economies
- ▶ Externalities
- ▶ Immigration and the City
- ▶ Internal Migration
- ▶ International Migration
- ▶ International Policy Coordination
- ▶ Public Goods
- ▶ Utilitarianism and Economic Theory
- ▶ Voluntary Contribution Model of Public Goods

Acknowledgment I received useful comments and helpful suggestions from two anonymous referees. All remaining errors are mine.

Bibliography

Andreoni, J. 1989. Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* 97: 1447–1458.

Bubb, R., M. Kremer, and D.I. Levine. 2011. The economics of international refugee law. *Journal of Legal Studies* 40: 367–404.

Cornes, R., and T. Sandler. 1996. *The theory of externalities, public goods and club goods*. 2nd ed. Cambridge: Cambridge University Press.

Czaika, M. 2009. Asylum cooperation among asymmetric countries: The case of the European Union. *European Union Politics* 10: 89–113.

Facchini, G., O. Lorz, and G. Willmann. 2006. Asylum seekers in Europe: The warm glow of a hot potato. *Journal of Population Economics* 19: 411–430.

Fernández-Huertas Moraga, J., and H. Rapoport. 2014. Tradable immigration quotas. *Journal of Public Economics* 115: 94–108.

Fernández-Huertas Moraga, J., and H. Rapoport. 2015. Tradable refugee-admission quotas and EU asylum policy. *CESifo Economic Studies* 61: 638–672.

Foged, M., and G. Peri. 2016. Immigrants' effect on native workers: New analysis on longitudinal data. *American Economic Journal: Applied Economics,* forthcoming.

Giordani, P.E., and M. Ruta. 2013. Coordination failures in immigration policy. *Journal of International Economics* 89: 55–67.

Hathaway, J.C., and R.A. Neve. 1997. Making international refugee law relevant again: A proposal for collectivized and solution-oriented protection. *Harvard Human Rights Journal* 10: 115–211.

Hatton, T.J. 2004. Seeking asylum in Europe. *Economic Policy* 19: 5–62.

Monheim-Helstroffer, J., and M. Obidzinski. 2010. Optimal discretion in asylum lawmaking. *International Review of Law and Economics* 30: 86–97.

Roth, A.E. 1985. The college admissions problem is not equivalent to the marriage problem. *Journal of Economic Theory* 36: 277–288.

Schuck, P. 1997. Refugee burden sharing: A modest proposal. *Yale Journal of International Law* 22: 243–297.

Suhrke, A. 1998. Burden-sharing during refugee emergencies: The logic of collective versus national action. *Journal of Refugee Studies* 11: 396–415.

Suriyakumaran, A., and Y. Tamura. 2016. Asylum provision: A review of economic theories. *International Migration,* forthcoming.

International Coordination of Regulation

David Lazer

Abstract

Trade among states with diverse regulatory systems creates the possibility of taking advantage of the cost differentials in production that result. However, it is increasingly clear that

what happens in one jurisdiction affects policy in other jurisdictions. It is often argued that this creates a 'race to the bottom' effect, where the most lax regulation gains an advantage, but the evidence on this is mixed, at best, and there is a plausible argument too for a 'race to the top' effect, where states set high regulatory standards as a barrier to entry.

Keywords

Comparative advantage; Delaware effect; California effect; Policy information interdependence; Race to the bottom; Race to the top; Regulation; Regulatory competition

JEL Classifications

K2; K23; L50; L51

Regulation is a pervasive element of modern government. The regulatory state has its hand in everything from the food we eat to the couches we sit on. The regulations we live with are necessarily a collective affair – as a general proposition two people who live in the same jurisdiction cannot choose different regulatory regimes. However, it is increasingly clear that regulation is a collective affair more broadly – what happens in one jurisdiction affects other jurisdictions. This interdependence creates potential governance challenges, in part due to strategic dilemmas that these interdependencies may create, as well as accountability issues that occur when important policies affecting a polity originate outside of that polity.

Samuelson (1949) offered an evocative thought experiment that sheds some light on the potential dilemmas. Imagine, Samuelson asked, the integrated economy of a world with no boundaries, no transportation costs. Now assume angels come along and divide the world up into different nations, with different allocations of factors of production. The question Samuelson posed was how international trade might allow the world to recapture that lost paradise, where the trade of goods is a substitute of sorts for a homogeneous distribution of factors of production.

Interestingly, international trade creates an opportunity where none existed in Samuelson's integrated economy, if we assume that Samuelson's integrated economy had a single regulatory regime. Specifically, trade among states with diverse regulatory systems creates the possibility of taking advantage of the cost differentials in production resulting from that regulatory heterogeneity. For example, if we imagine that state A has strict regulation of public good 1 and relaxed regulation of public good 2, and state B has relaxed regulation of public good 1 and strict regulation of public good 2, then both states can benefit from trade. Those sectors that can produce more cheaply in A (because of its more relaxed regulation of public good 1) will naturally arise there (even in the absence of capital mobility), and similarly with respect to sectors that would produce more cheaply in B. Generally, the basis for trade (national and international) rests in significant part on the pillar of heterogeneity (that is, comparative advantage). Samuelson's angels create the possibility of exchange based on regulatory heterogeneity that did not exist in the integrated economy.

The preceding assumes that policy is exogenous and that all factors of production are immobile. What happens if we relax these assumptions? One possibility is that regulatory policies will diverge, because trade reduces the adverse economic effects of those policies. To take an extreme example, imagine a state with a preference for strong regulation of a particular sector that makes some good for which there is inelastic demand. In a closed economy the benefits of strict regulation need to be weighed against the fact that the costs of that regulation will be fully borne by consumers. In an open economy, it may be possible to regulate that sector out of existence, with fairly minimal welfare impacts on society – because that sector may locate in another jurisdiction where the demand for regulation in that sector is lower. That is, it is theoretically possible that trade will enable some jurisdictions to regulate more strictly, thus raising the average level of regulation in the international system.

Receiving far more attention, however, is the possibility of a 'race to the bottom' (RTB), where

the combination of trade and factor mobility yields uniform downward pressure on regulation across all jurisdictions. RTB has been asserted in many settings, although proven in few. The essential intuition is fairly simple. Consider the following simple model of the world, where there are two factors of production, say labour and capital. Labour is assumed to be immobile and capital mobile. If we assume that labour and capital are complements in production at the national level, this would yield a competition at the national level for division of the surplus produced through their combination. This intranational competition, in turn, has an international dimension because capital can flow to jurisdictions that offer the biggest share of that surplus. That is, effectively, labour from different jurisdictions will compete to attract capital. More capital will yield a larger surplus; however, the competition among jurisdictions means that most of that surplus will go to capital. We can see this type of dynamic, for example, in the competition among states in the United States to attract automobile factories, where, as Donahue (1997) documents, foreign manufacturers garnered enormous incentives from US state governments.

There is also clearly a RTB effect when there are physical externalities from one jurisdiction to another – for instance, when power plants are placed close to borders, and pollution spills over to the neighbouring jurisdiction. Except where noted, for this paper the case of physical externalities will be bracketed, because analytically it is uninteresting. In other words, it is clear that there is a potential for a collective failure when there are physical externalities (the most obvious contemporary example is carbon emissions).

When applied to the regulatory context, regulation may be seen, in part, as an effort by a jurisdiction to reallocate some of the surplus towards immobile actors. Environmental protection offers a nice illustration. Efforts to protect the local environment increase costs to capital. As a result, capital is less likely to locate in a highly regulated jurisdiction, lowering wages (relative to a more lax regulatory jurisdiction) to the point that capital is indifferent between locating in jurisdictions with different levels of regulatory stringency. The preceding discussion would suggest that given a large number of jurisdictions, and a lack of collusion among those jurisdictions, the net effect of dividing the world up will be to redistribute from immobile factors of production to mobile factors of production. In the example above, this redistribution would take place from the environment and from labour to capital.

The above discussion notwithstanding, RTB has had far more currency among politicians than economists. While analytic models, following from the studies of regulatory federalism by Tiebout (1956), sometimes find the possibility of an RTB, it is far from the typical finding (Oates 2002). It should be noted, however, that the federal context is an imperfect analogue for the international context. For example, these models assume strong sorting effects of citizens with respect to, for example, preferences for environmental protection. It is less likely that such significant sorting effects exist at the international level.

Similarly, empirically, there is relatively little support for RTB dynamics. For example, there is little evidence that firms locate based on relaxed regulatory regimes (Bartik 1988). There is also little evidence that increased capital mobility over the last few decades has created downward movement in regulation, and some evidence that supports the opposite proposition (such as Engel 1997; Fredriksson and Millimet 2002; Frankel and Rose 2005).

The lack of RTB effects may in part reflect that costs imposed by a regulatory system are a fairly small factor in the decision on location (Jaffe et al. 1995). Further, one likely and important reason for the lack of observation of RTB dynamics is the inability of researchers to capture the full range of reasons why firms decide to locate in particular jurisdictions. The decision to locate an oil rig in a particular location may be affected by environmental regulations, but first and foremost is certainly driven mostly by whether there is oil present at that location. An oil company, all else being equal, may prefer to drill in locations with less stringent environmental rules. However, those locations with the most desirable locations for drilling oil are also therefore in a position to seek a larger share of the income produced from that oil. This share, in those jurisdictions with preferences for a cleaner

environment, would likely in part be extracted through stronger environmental regulations. Empirically, this might yield the outcome that those locations with the most drilling will also have the strongest regulations. Such a snapshot might be misleading, since it might still be the case that regulatory competition results in more lax regulation than in a counterfactual world where capital were not mobile, or where policymakers colluded.

More generally, there are a variety of reasons why particular locations might offer particular advantages for capital. Some of these may be exogenous (such as those related to the location of particular natural resources). Others may be endogenous. For example locating particular physical capital (such as factories) in a specific location might facilitate the creation of human capital in that location, which might be relatively immobile. Closely related, there might be returns to scale at the industry-jurisdiction level. That is, the productivity of a firm might be positively related to the number of other firms in a jurisdiction or region. In fact, clustering of sectors in particular regions is quite common (Krugman 1996). If the emergence of such a cluster reflects external economies, this creates the possibility that a jurisdiction containing such a cluster can increase the stringency of its regulations without a worry that capital will flee (Baldwin and Krugman 2004).

The debate in the legal literature regarding the clustering of industry incorporations in Delaware offers an illuminating case study that illustrates this point. The starting point for this literature is the observation that Delaware towers over the rest of the United States in terms of number of incorporations. This was originally viewed as evidence of a lax regulatory regime in Delaware (Cary 1974). The essential argument was that corporations located in Delaware because doing so minimized the burdens placed on corporate management, often at the expense of shareholders. The hypothesized RTB was branded, as a result, as the 'Delaware effect'. This argument came under sharp criticism in the 1990s. The first critique was that there would be shareholder pressure (exerted in part through the stock price) against locating in a jurisdiction that did not preserve shareholder value (Revesz 1992). The dominance of Delaware thus reflects the capacity of Delaware to effectively manage corporate governance while still preserving shareholder value. That is, the dominance of Delaware reflects the benefits of regulatory competition, where Delaware simply provides superior governance to everyone else. A second critique (of both the benefits and the costs of regulatory competition) was that there are economies of scale in providing good corporate governance (Kahan and Kamar 2003). For example, good corporate governance requires predictability, and predictability is facilitated by ample case law. Delaware, by dint of historical accident, had garnered an insurmountable lead in effectively producing good corporate law. The implication of this, in turn, is that Delaware has a fair degree of slack in extracting some surpluses from the regulated parties, as long as it does so in a way that does not threaten its competitive advantage in corporate governance.

Note that there are RTB debates in other domains, such as welfare benefits (Dahlberg and Edmark 2008; Bailey and Rom 2004) and corporate tax rates (Basinger and Hallerberg 2004).

While these RTB debates rage on, with strong intuitions and political appeal ranged against modest analytic and empirical support from the economics literature, there is a stronger consensus about the potential capture of the regulatory system by domestic interests (Bartel and Thomas 1987) – which I will label 'regulation as protection' (RAP). To paraphrase Clausewitz, regulation may be viewed as market competition through other means. It is rare for regulation to be neutral in its impact on producers in a given sector. A more restrictive regulation has the potential to benefit some producers at the expense of others. For example, requirements for greater efficiency in cars benefit producers that already produce high-efficiency cars, at the expense of producers of low-efficiency cars. In the context of international trade, the alignment of interests will often be domestic producers opposed to international producers interested in entering the domestic markets and domestic consumers who might benefit from increased competition in the home market.

The potential scenario is for domestic producers to hijack the domestic regulatory apparatus as a means to block international competition.

It is also clear that in many sectors there are strong drivers for convergence in regulation. A simple example is the mandate in a jurisdiction to drive on one side or the other of the road. Given significant traffic between two jurisdictions, it is clear that there would be a strong benefit to a consistent standard for driving. Thus, for example, Canada, which had a patchwork of standards in the first half of the 20th century at the provincial level, with some provinces mandating left-hand traffic, and others right-hand traffic, gradually converged to right-hand traffic, with uniformity emerging by the middle of the century. This presumably was driven in part by the need for internal consistency, and in part because of the degree of traffic between the United States and Canada. In the regulatory context, such convergence may be seen fairly routinely across policy areas, for example in food safety (Lazer 2001), where small countries adopt the standards of their large export markets.

Regulatory convergence more generally may take place to guard export markets. As Vogel (1995) argues, this type of push for convergence often occurs in international trade when there are increasing returns to scale in production, combined with potential divergence in product standards. Thus, for example, Vogel documents the flow of environmental standards for automobiles around the world, which he labels the race to the top (RTT) 'California effect' in contrast to the RTB 'Delaware effect'. Vogel argues that there is potentially a systematic bias toward higher standards because of the efficiency imperatives of having consistent standards. Such an imperative should lead toward convergence on the strictest standard on the system (or at least a standard that would encompass the large majority of the system), so as to achieve efficiencies in scale of production. Such a diffusion pattern should follow the reverse direction of exports, for example, as Prakash and Potoski (2006) found with respect to the spread of ISO 14001 adoption.

Finally, even in the absence of trade or factor mobility, there is significant potential for regulatory interdependence due to policy information interdependence (PII). Policy is necessarily experimentation, and novel policies even more so. Regulation in jurisdiction A creates insights in jurisdiction B as to what would be good or bad policy, where this information spreads through various informational networks (Wolman and Page 2002; Lazer 2005). Emulation may reflect lesson drawing or serve the function of policy legitimation (Bennett 1997; Busch et al. 2005). There is a substantial literature on policy emulation (e.g. Rose 1993; Haas 1992), and it is clear that there is no reason to expect domestic regulatory policy to be exempt from emulation (e.g. Simmons and Elkins 2004).

This array of interdependencies offers an array of empirical challenges for the researcher and governance challenges for the policy maker. For the researcher, the potential presence of different processes (which, if validated, have very different normative implications) creates a difficult but not impossible nut to crack (for various recent efforts to deal with exactly this issue, see Braun and Gilardi 2006; Simmons et al. 2006; Levi-Faur 2005).

For the policy maker each of these kinds of regulatory interdependencies creates different types of collective strategic interaction problems (Lazer 2001, 2006), which in turn translate into a collective governance challenge (Scharpf 1997). RTB and RAP might be categorized as a prisoner's dilemma, where the potential dysfunctional equilibrium would be either suboptimally lax standards in the case of RTB, or suboptimally strict standards in the case of RAP. RTT may be viewed as a coordination game, where one potentially problematic outcome is the emergence of suboptimally strict standards (or perhaps worse would be the case where the RTT did not occur, where a critical mass of support did not emerge for any standard), with a handful of large jurisdictions driving the standards for the world. All of these issues might call for some type of negotiated standard. However, such centrally negotiated standards might eliminate some of the very benefits of trade in the first place. And the presence of PII would suggest a different kind of public goods issue than is usually associated with regulation – the public good of information.

This construction of regulatory interdependence suggests a dual conundrum. The first aspect of this is how to take advantage of the publicness of the information, and the second is how to support continued production of this public good (Lazer 2005).

See Also

► Comparative Advantage

Acknowledgment Thanks go to Elta Smith, for her research assistance with respect to this paper.

Bibliography

Bailey, M., and M.C. Rom. 2004. A wider race? Interstate competition across health and welfare programs. *Journal of Politics* 66: 326–347.

Baldwin, R., and P. Krugman. 2004. Agglomeration, integration and tax harmonization. *European Economic Review* 48: 1–23.

Bartel, A., and L.G. Thomas. 1987. Predation through regulation: The wage and profit effects of the Occupations Safety and Health Administration and the Environmental Protection Agency. *Journal of Law and Economics* 30: 239–264.

Bartik, T. 1988. The effects of environmental regulation on business location in the United States. *Growth and Change* 19: 22–44.

Basinger, S., and M. Hallerberg. 2004. Remodeling the competition for capital: How domestic politics erases the race to the bottom. *American Political Science Review* 98: 261–276.

Bennett, C. 1997. Understanding ripple effects: The cross-national adoption of policy instruments for bureaucratic accountability. *Governance* 10: 213–233.

Braun, D., and F. Gilardi. 2006. Taking 'Galton's problem' seriously: Towards a theory of policy diffusion. *Journal of Theoretical Politics* 188: 298–322.

Busch, P., H. Jorgens, and K. Tews. 2005. The global diffusion of regulatory instruments: The making of a new international environmental regime. *Annals of the American Academy of Political and Social Science* 598: 146–167.

Cary, W. 1974. Federalism and corporate law: Reflections upon Delaware. *Yale Law Journal* 83: 663–666.

Dahlberg, M., and K. Edmark. 2008. Is there a 'race-to-the-bottom' in the setting of welfare benefit levels? Evidence from a policy intervention. *Journal of Public Economics* 92: 1193–1209.

Donahue, J. 1997. *Disunited states*. New York: Basic Books.

Engel, K. 1997. State environmental standard-setting: Is there race and is it to the bottom? *Hastings Law Journal* 48: 271–398.

Frankel, J., and A. Rose. 2005. Is trade good or bad for the environment? Sorting out the causality. *Review of Economics and Statistics* 87: 85–91.

Fredriksson, P., and D. Millimet. 2002. Strategic interaction and the determination of environmental policy across U.S. States. *Journal of Urban Economics* 51: 101–122.

Haas, P. 1992. Introduction: Epistemic communities and international policy coordination. *International Organization* 46: 1–35.

Jaffe, A., S.R. Peterson, P.R. Portney, and R.N. Stavins. 1995. Environmental regulations and the competitiveness of U.S. manufacturing: what does the evidence tell us? *Journal of Economic Literature* 33: 132–163.

Kahan, M., and E. Kamar. 2003. The myth of state competition in corporate law. *Stanford Law Review* 55: 679–750.

Krugman, P. 1996. *The self organizing economy*. Cambridge, MA: Blackwell.

Lazer, D. 2001. Regulatory interdependence and international governance. *Journal of European Public Policy* 8: 474–492.

Lazer, D. 2005. Regulatory capitalism as a networked order: The international system as an informational network. In *The rise of regulatory capitalism: The global diffusion of a new order*, ed. D. Levi-Faur and J. Jordana. London: Sage.

Lazer, D. 2006. Global and domestic interdependence: Modes of interdependence in regulatory policymaking. *European Law Journal* 12: 455–468.

Levi-Faur, D. 2005. The political economy of legal globalization: Juridification, adversarial legalism, and responsive regulation. A comment. *International Organization* 59: 451–462.

Oates, W. 2002. Fiscal and regulatory competition: Theory and evidence. *Perspektiven der Wirtschaftspolitik* 3: 377–390.

Prakash, A., and M. Potoski. 2006. Racing to the bottom? Trade, environmental governance, and ISO 14001. *American Journal of Political Science* 50: 350–364.

Revesz, R. 1992. Rehabilitating interstate competition: Rethinking the 'race to the bottom' rationale for federal environmental regulation. *New York University Law Review* 67: 1210–1254.

Rose, R. 1993. *Lesson-drawing in public policy*. Chatham: Chatham House.

Samuelson, P. 1949. International factor price equalization once again. *Economic Journal* 58: 163–184.

Scharpf, F. 1997. Introduction: The problem-solving capacity of multi-level governance. *Journal of European Public Policy* 4: 520–538.

Simmons, B., and Z. Elkins. 2004. The globalization of liberalization: Policy diffusion in the international political economy. *American Political Science Review* 98: 171–189.

Simmons, B., F. Dobbin, and G. Garrett. 2006. Introduction: The international diffusion of liberalism. *International Organization* 60: 781–810.

Tiebout, C. 1956. A pure theory of local public expenditures. *Journal of Political Economy* 64: 416–424.

Vogel, D. 1995. *Trading up: Consumer and environmental regulation in a global economy.* Cambridge, MA: Harvard University Press.
Wolman, H., and E. Page. 2002. Policy transfer among local governments: An information-theory approach. *Governance* 15: 477–501.

International Economics, History of

M. June Flanders

Abstract

Starting with mercantilist theories, the article deals with laissez-faire rejections of mercantilism, the Ricardian justification of free trade and its extension to multiple countries and commodities. Heckscher–Ohlin trade theory, factor-price equalization and the 'Leontieff paradox' debate follow. Intra-industry trade is related to increasing returns, imperfect competition, and product differentiation. Trade and growth, economic geography, and tariffs and trade restrictions are summarized. Regarding macroeconomics, Hume and the monetary approach to the balance of payments are compared with income adjustment theories. International monetary regimes, exchange rate regimes, capital transfers, internal–external balance, and new international macroeconomics are discussed.

Keywords

Absorption approach to the balance of payments; Balance of trade; Balance of payments; Bank of England; Beggar-my-neighbour; Brand identification; Cantillon, R.; Classical economics; Comparative advantage; Cone of diversification; Convertibility; Corn Laws; Devaluation; Diffusion of technology; Dirigisme; Economic geography; Elasticities approach to the balance of payments; Eurozone; Exchange rate regimes; Factor price equalization; Fixed exchange rates; Fleming, J. M.; Flexible exchange rates; Foreign direct investment; Full employment; Gains from trade; Gervaise, I.; Gold standard; Growth and international trade; Haberler, G.; Hayek, F.; Heckscher–Ohlin trade theory; Heckscher–Ohlin–Samuelson model; Hume, D.; Imperfect competition; Increasing returns; Infant-industry protection; Innovation; Intermediate products; International capital flows; International economics, history of; International monetary theory; International trade theory; Internal–external balance; Intra-industry trade; Labour theory of value; Leontief paradox; Lerner, A. P.; Locke, J.; Long-term economic growth; Meade, J. E.; Mercantilism; Microfoundations; Monetary approach to the balance of payments; Money multiplier; Money supply; Multinational corporations; Mundell, R.; New classical macroeconomics; New international macroeconomics; Offer curve or reciprocal demand curve; Ohlin, B.; Outsourcing; Product differentiation; Production possibility frontier; Quantity theory of money; Quotas and tariffs; Ricardo, D.; Smith, A.; Specialization; Specie-flow mechanism; Sterilization; Sticky wages; Subsistence; Taussig, F.; Technological progress; Terms of trade; Tobin tax; Torrens, R.; Trade policy, political economy of; Transfer problem; Vent for surplus; Walrasian general equilibrium

JEL Classifications

B2

The Real Theory of International Trade

Contemporary international trade theory has its roots in classical economics, which developed in opposition to the widely accepted views, known (since Adam Smith's introduction of the term into British discourse) as mercantilism, held until the mid- 18th century by both policymakers and many analysts. This was a loose body of doctrine advocating extensive government control and interference in economic activity. In the context of international trade it refers to the imposition of

tariffs, quotas, and prohibitions, designed to maximize the balance of payments surplus and the net inflow of precious metals (specie). The justification for such policies took many forms. Since what we know as mercantilism lasted for about two centuries and arose in widely differing social and political contexts, there was room for substantial differences of opinion. In some cases it was apparently a fear of goods, in others an identification of specie with real wealth, and in still others, as Keynes (1937) suggested, a desire to stimulate employment. (Detailed critical surveys and analyses are to be found in Heckscher 1935; Viner 1937; Blaug 1985.)

Though he was not the first to oppose dirigisme and to see the advantages of trade and specialization, Smith, in his *Wealth of Nations* (1776), provided the starting point for classical theories of trade. He argued both that gold and silver are not real wealth and that generating a balance of trade surplus is not the only way to acquire them. His discussion of interference with trade in goods and services is the same as his treatment of other governmental interferences: he dealt with the interactions of markets for goods and factors, showing the effects of tariffs and subsidies on each. He recognized situations in which restrictions on trade might be justified: these included defence needs, retaliation and infant-industry arguments.

Smith did not develop a theory of comparative advantage, though he came close to it when he noted that Britain was more productive in manufactures relative to Poland than it was in agriculture (Smith 1776, pp. 6–7). But his main argument in favour of trade is indeed that which is now labelled the 'vent for surplus'. This is enunciated at several points, but a clear statement occurs when he says that, because of international trade, 'the narrowness of the home market does not hinder the division of labour in any particular branch of art or manufacture from being carried to the highest perfection. 'By opening a more extensive market for whatever part of the produce of their labour may exceed the home consumption' (1776, p. 415). Mill (1848) cites Smith as having viewed exports as an outlet for surplus, and Bastable (1897) takes Smith very literally on the surplus and attacks him vigorously for it, arguing that it implies the existence of unemployed resources. Schumpeter (1954, p. 374) accuses Smith of having 'believed that under free trade all goods would be produced where their absolute costs in terms of labor are lowest'. But Schumpeter notes that Viner indicates that Smith, and others before him, had formulated the more general proposition that, under free trade, commodities would be imported whenever they could be obtained most cheaply in this way. This includes the case where exports 'cost less to produce than it would cost to produce the corresponding imports at home and thus implies the theorem of comparative costs'. A contemporary evaluation of the several interpretations of Smith's trade theory is provided by Blecker (1997).

It was Ricardo, who, in his *Principles of Political Economy and Taxation* (1817), first articulated explicitly and emphasized and publicized the theory of comparative advantage (though Torrens, in 1815, had come very close to it), noting that absolutely lower costs in the production of all goods was not a sufficient reason for producing them all, and that it was generally to the advantage of a country to specialize in the production of that which it did best. Here, too, the objective was to influence policy. Ricardo, like Smith, developed his views on trade within the framework of his model of how an economy did or should operate. British manufacturers had no need of protection from imports, and his main concern was with duties on grain imports, the Corn Laws, which protected British agriculture. Ricardo attacked these because they raised food prices and hence real wages at the expense of profits, thus discouraging investment and growth. The equilibrium result would be zero saving and investment and a long-run stationary economy with wages at subsistence level. With free trade, national income would be maximized by keeping real wages in industry low and by concentrating output in the relatively low- cost sectors, regardless of absolute advantage or disadvantage. It was here that he inserted a brief monetary statement, that there was a natural distribution of specie, which depended on the economic size of each

country and the parameters of its monetary variables.

Ricardo's trade theory, like classical economics in general, is based on a model of supply: price is the long-run supply price. He implies that complete specialization is the norm, but he neither makes this explicit nor specifies how the gains from trade are divided between the trading countries. He considered the possibility of non-traded goods and the idea that a country could produce and export more than one good if country size or demand patterns were highly uneven.

Mill (1848, pp. 583–606) brought demand and, implicitly, its elasticity into the theory, thereby explaining how the terms of trade were established between the limits set by comparative advantage. He extended the analysis to include transport costs, more commodities (which limited the range of the terms of trade) and more countries. He also noted that, since the terms of trade must be the same for all countries, the gains from trade will be greater for those for which the opportunity cost of the exportable good is lower. These refinements were spelled out precisely in the 20th century by Graham (1923), who noted that the terms of trade and the probability of specialization depended on the relative size of the (two) trading countries and the relative importance of the (two) traded goods in total consumption. Furthermore, when more countries and/or more goods were brought into the picture, the final terms of trade were narrowed down even further, as was the exact number of goods traded by any given country.

Marshall (1879; 1923) generalized the theory into a two-country multi-commodity analysis (using the device of 'bales' of goods) and derived offer curves to depict graphically the general equilibrium in production and consumption that Mill had analysed only verbally. Edgeworth (1925) produced a similar analysis. The derivation of the offer (reciprocal demand) curves was not spelled out; all domestic markets were assumed to be equilibrium at each point on an offer curve. Marshall (1879) has a detailed analysis of equilibrium and stability conditions, but not, explicitly, elasticity.

A sizeable literature developed, well into the 20th century, both testing Ricardian comparative cost theories and justifying economists' generally free-trade position. In a fairly laboured list of possibilities, Samuelson (1939) showed that some trade is always better than no trade. This holds for various shapes of production possibilities curves, but its welfare implications depend on the welfare device of the ability of gainers to compensate the losers, or, as he put it, 'by Utopian co-operation everyone can be made better off as a result of trade' (1939, p. 204).

In the 1930s and 1940s, there was extensive discussion of arguments for and against tariffs, quotas, and other impediments, and the real gain – or loss – from imposing them. This included the debate on the relationship of trade structure to growth, which continues to the present.

A major paradigm shift came with the formal use of what are often called 'neoclassical' assumptions about technology and preferences in the analysis of comparative costs. In the entry in this dictionary on Haberler, Gottfried, it is claimed that, although Barone in 1908 (but not subsequently) had a (non-concave) production-possibility frontier and a community indifference curve, it was Haberler's independent discovery in 1930, and the use to which he put it, that transformed the theory of international trade. Haberler thereby broke with the labour theory of value, the production possibility frontier becoming standard in all economic theorizing and teaching. Lerner went on to draw a 'compound indifference curve' in 1932, and in 1934 developed the demand side fully. Both Lerner and Haberler (1936) dealt with the possibility of increasing returns, Haberler granting that this could justify tariffs.

In fact, Haberler was preceded by Bickerdike, and by Heckscher in 1918 and Ohlin in 1924, but the latter two published in Swedish, reducing the visibility of their work. They addressed the issue of factor prices under free trade. Ohlin published an enlarged version of his 1924 Ph.D. thesis in English in 1933; Heckscher's seminal paper was translated (partially) into English only in 1949.

Heckscher's general equilibrium analysis, in 1918, was wholly verbal. He examined the

reasons for, and results of, large-scale Swedish migration to America. In studying this, he showed that under certain circumstances trade in goods could substitute for movement of factors in equalizing factor prices. If factor endowments were not too different, factor prices would inevitably be equal throughout the world. Ohlin, his student, drew from this and, using Cassell's version of Walrasian general equilibrium analysis, developed a more formal approach (1924). Lerner demonstrated factor price equalization with arithmetic and geometry in a seminar paper in 1933, the Swedish work being unknown.

Ohlin himself rejected the conclusion of the equalization of factor prices. In his formal analysis he simply (inexplicably) assumed that each country was completely specialized; in verbal discussion, he always saw exceptions to any generalization. Among the obstacles to factor price equalization were insufficient geographical and occupational mobility of factors between industries within a country, increasing returns to scale, excessive imbalance in factor supplies, taxes, transport costs, and imperfect competition. All these had to be taken into account when explaining the actual patterns of trade. Furthermore, the verbal analysis is explicitly dynamic. One example is his discussion of the effects, almost year by year, of an increase in world demand for a country's exports. There is an increased demand for labour in that industry, so labour moves into it; individuals move, gradually, into the skill group required for those goods; labour moves, again gradually, geographically. All these take time: some years later, if demand remains as it was with no further increases, the supply of productive factors will have adjusted to the new demand levels. Ohlin's earlier work (his thesis) is replete with numerical examples of international and interregional differences in price levels, in expenditures for food and housing, and the like.

The factor price equalization aspect of the model was largely ignored until Stolper and Samuelson unearthed it in 1941. (When White reviewed Ohlin's book in 1934 and compared his model with the first German edition of Haberler's *International Trade*, he did not mention the factor price equalization issue at all.) The model was formalized subsequently by Samuelson (1948, 1949, 1951–2, 1953–4), and later elaborated, extended, and modified by Jones (2000, passim) and others. Originally these formalizations required very strict assumptions: two commodities, two countries, constant returns to scale, both goods being produced in both countries. The last requirement of the Heckscher–Ohlin–Samuelson model, as it became known, was very stringent. It meant that factor proportions had to be within what was called the cone of diversification, where both goods could be produced when prices were equalized across countries. Later models picked up some of the complications that Ohlin had enumerated earlier: Jones developed variations, and welfare implications, for the cases of specific factors (immobile between industries), increased numbers of goods and factors, trade in intermediate goods, and tariffs. A penetrating discussion of fundamental methodological issues in the development of the theory and the subsequent debate on the Leontief paradox is provided by De Marchi (1976).

The Heckscher–Ohlin paradigm remained untested and empirical work in trade proceeded along Ricardian lines until Leontief applied his input–output model to a test of United States trade and concluded that US exports were labour-intensive relative to imports. The implied paradox gave rise to an avalanche of theoretical and empirical work, ranging from multifactor models to specific explanations in terms of the period and the structure of world trade in the early post-Second World War period which was used in the study. One explanation offered was the possibility of reversals of relative factor intensities as relative factor prices changed, a proposition which led to a spate of ongoing debates in capital theory (Arrow et al. 1961.)

Baggott (1970) presents an exhaustive list and discussion of the various proffered explanations, including her own, that the United States in the relevant period was exporting capital directly through its balance of trade surplus, and also notes the possibility that some of the capital-intensive commodity imports were produced by branches of American firms, with American

capital. Caves and Jones (1977) give some of the highlights in the debate. Others sought to resolve the paradox in a wide variety of ways: for example, by claiming that there had been a conceptual misunderstanding, that labour needed to be augmented by considering human capital, and that during Leontief's sample period capital intensity was highly correlated with high natural-resource use. For an exhaustive treatment of the paradox and of the numerous attempts to resolve it, see Chipman (1965–66, 33 51–70).

Economic analysis and investigation move with events. Both output and trade became increasingly characterized by differentiated products, brand identifications, oligopolistic behaviour and the scale economies which these generated, strategic investment and marketing decisions, and expanding trade in intermediate products. Historically, the first phenomenon observed in the post-Second World War period was the large and growing intra-industry trade (see Grubel and Lloyd 1975). Helpman and Krugman (1985) treat such trade as a result of product differentiation, generally associated with monopolistic competition and increasing returns, and model it as coexisting with inter-industry trade based on factor endowments. A survey of trade theory based on imperfect competition due primarily to increasing returns can be found in Krugman (1987). The behaviour of multinational corporations and conglomerates, often producing and marketing a wide variety of products, demanded examination and explanation (see Caves 1982). Dixit and Norman (1980) include in their formal, general-equilibrium rewriting of trade theory the case of oligopolistic markets. Game-theoretic study of innovation strategies, outsourcing, and the political economy of trade restrictions policy followed. Refinements of these trends continue and are at the forefront of international trade theoretical and empirical work today (such as that of Grossman and Helpman 1991, 2002).

There has also developed a large literature on the mutual interaction of growth and international trade, which is a subject in and of itself, starting with Smith's vent for surplus theory, through issues of putative exploitation of developing countries, into contemporary empirical/historical studies of long-term growth, and the effects of differential increases in the several factors of production. The differential rates of technical progress, the diffusion (or lack of it) of technology, and the concomitant international differences in growth rates are some of the topics being analysed and explored. The existence of increasing returns, both to firms and to industries, first discussed by Haberler, has led to renewed interest in economic geography, a field being examined today both by economists and geographers and, perhaps, waiting for some real interdisciplinary exploration. A critical review of the field, and a plea for integration of new explorations in economic geography with industrial geography is offered by Martin and Sunley (1996).

Macro-Monetary Theory

Many of the mercantilists, with their concern for the accumulation of specie, evinced no sense of the impossibility of having a permanent balance of trade surplus. Some saw money as working capital that would drive an increasing volume of trade; others saw little problem in absorbing specie, perceiving trade as constrained by a shortage of coin. In the 16th and 17th centuries, there was increasing awareness of a link between money and the price level, culminating in Locke's formulation of the quantity theory (1696). Despite a partial anticipation of the result by Cantillon (1755; written c. 1730), it is Hume (1752) who is commonly credited with the price-specie- flow mechanism and the implied endogeneity of the money supply in an open economy. However, both were preceded by Isaac Gervaise, whose pamphlet of 1720 was almost totally ignored. Gervaise had an adjustment mechanism, essentially the monetary approach to the balance of payments (Gervaise 1720), which Ricardo 100 years later was to enunciate as the 'natural distribution of specie'. Beyond that, he had a model of financial adjustment through a money multiplier and real effects in the form of inter-industry shifts in production.

Hume argued that attempts to acquire precious metals, or prevent their export, would result in

price level changes, affecting the balance of trade and reversing the specie flows. The same line of reasoning (though under flexible exchange rates) informed the work of Ricardo (1811) and Henry Thornton (1802). Other highly sophisticated monetary writings of the time are discussed in Hollander (1910–11). The main point, that the money supply is endogenous, resurfaced almost 150 years later in the monetary approach to the balance of payments (see Frenkel and Johnson 1976).

In subsequent British writings, through much of the 19th century, the distinction between international and domestic monetary theory barely existed. The British economy was open; the London money market was the world's financial centre. The brilliant stream of monetary debate in 19th-century England (see Fetter 1965; Bagehot 1873), carried out largely by bankers and business people, concentrated on issues of monetary policy for an open economy, but with little attention to the effect of policy on the real sector or on long-term financial markets. The analysis focuses in general on the short run: international disturbances were both exogenous and temporary in that world, and the issue was really how to ride the storm – the underlying structure was always, implicitly, in equilibrium. This was a reasonably accurate picture of the England of the day, where long-term capital outflows were effected by changes in trade flows in the equilibrating direction, and much of any needed real adjustment was performed in the periphery (see Ford 1962). This literature had to do primarily with the reciprocal relationship between specie holdings and specie flows on the one hand and the domestic money supply on the other. Much of this involved differences in the definition of money, the role of the Bank of England, the vulnerability of the bank's gold stock, by law subject to drain by holders of Bank of England notes (paper money), and the appropriate measures which this very public, privately owned, institution could and should take to protect itself. (Evidence of the time, and subsequent histories, point to the uniqueness of the Bank of England; yet in the paradigm of the adjustment mechanism it is always treated as the prototype of central banks.)

Convertibility of Bank of England notes into gold was suspended during the Napoleonic Wars; when it was resumed, in 1819, the bank entered a period of more than half a century of recurrent crises and near crises, but managed to maintain the convertibility of its notes into gold. Discussion and debate in this environment was almost brought to an end when Walter Bagehot published his influential *Lombard Street* in 1873. Based on experience, and possibly as a result of his hectoring, the Bank of England learned in the subsequent decades how to handle its huge constituency of world and domestic finance on the basis of very small reserves of gold, developing, gradually, a number of highly sophisticated 'tricks' in the money markets to protect itself and forestall runs. Nothing about this management, meticulously documented by Sayers in his two separate studies, had to do with interactions between the real and financial sectors, certainly not in the context of long-run relationships. There was little if any discussion of the balance of trade, except when the original disturbance was a trade imbalance (usually temporary), such as a crop failure. And nothing remotely suggested automaticity of adjustment; the effects on the domestic money supply were to be avoided or at least mitigated. This relative neglect continued in discussion of British monetary policy and international adjustment until 1925, when it appeared in the debate on the resumption of gold convertibility. The previous examination, by the Cunliffe Committee in 1918, was noteworthy for its brevity. There was really nothing to discuss: specie outflow led to the Bank of England's changing bank rate, which reversed the outflow. Changes in expenditures play a secondary role in the process of adjustment to gold flows (see Flanders 1989).

Despite these objections, the automatic price-specie flow mechanism, stemming from Hume, lived on, until it was challenged in the 20th century by several analyses of what became known as the transfer problem. The origin of this is traced to Thornton, Ricardo, and others in the early 19th century who distinguished between money transfers and deficits caused by harvest failures. Bastable (1889) emphasized the impact of a monetary transfer, such as a tribute, on demand and the

possibility of effecting it with no change in the terms of trade. But this insight faded and reappeared only in the 20th century, when Taussig and his Harvard students, including Jacob Viner, John Williams and Harry Dexter White, examined the adjustments to the huge capital flows of the 19th and early 20th centuries. They found that income and expenditure changes played a much more critical role than had been thought. Adjustment was too smooth and too fast for it to have worked through Humean kinds of changes in price levels and thence in trade balances.

It was at this time, during the academic year 1922–23, that Ohlin, visiting Harvard, developed the approach first expressed in his 1929 debate with Keynes over German reparations, and then in great detail in his book (1933). He spelled out explicitly an expenditure-driven adjustment to international capital movements. (The irony is that Keynes could not see this.) Capital flows would lead to changes in total spending and hence directly in net exports. Price and wage changes might ensue but were not essential to the adjustment process.

The price-specie-flow story was challenged later, from a different flank, by Brown (1940), who demonstrated that the canonical view of the historical gold standard was mistaken and that in fact it had been a sterling standard, managed by the Bank of England. The system worked reasonably smoothly because trade and long-term capital movements were consistent with long-run equilibrium in the balance of payments. When this ceased to be true, England went off gold. Following the chaos of the inter-war period and the controls of the war and post-war years, the establishment of the International Monetary Fund constituted a recognition that the textbook adjustment mechanism of a metallic standard could not be relied upon. But what emerged in fact was a dollar standard rather than the intended multilateral system.

The textbooks continued to describe the international monetary system in terms of the price-specie flow mechanism and to treat capital movements as either factor flows (foreign investment) or short-term financial adjustments. In 1937, Hayek had taken the position that there had never been a true test of the price-specie flow mechanism in a multilateral world system in which domestic money supplies were endogenous and adjustments were automatic. In a neglected lecture at the London School of Economics he argued that the fixed exchange rate system, or gold standard, should not be abandoned on grounds of its failure, since it had never been operated correctly. There had never been a time, he said, when domestic money supplies were made to vary in response to specie flows as they would have had there been no sterilization or offsetting policies, that is, had the specie-flow mechanism functioned in the manner of the traditional paradigm.

Hayek's complaint, in the mid-1930s, was made in response to growing sentiment in favour of fluctuating or, more accurately, administratively pegged exchange rates. Given the willingness to consider changing the peg, the rate became a policy tool and a literature developed around what was called 'internal–external policy'. Beginning with Joan Robinson's attack on 'beggar-my-neighbour' devaluations (1937), there ensued a discussion of the effects of exchange rate changes on income and expenditure, as well as on the balance of payments; starting with the elasticities approach (elasticities of demand for and supply of exports, which proved to depend on general equilibrium in the domestic goods markets), moving on to the absorption approach to the balance of payments (introducing monetary effects as devaluation altered price levels) and culminating in Meade's massive multi-equation model of an open economy (1951). Meade rang the changes of the effects of various domestic monetary and fiscal policies directed at the level of employment and balance of payments equilibrium, under different conditions of price flexibility, capital mobility, wages policy, and various types of initial disturbance in regimes of (*a*) pegged and (*b*) flexible exchange rates (for a more detailed account, see Flanders 1989). Not dissimilar in aim and scope is a neglected attempt by Stuvel (1950) to formalize the effects of exchange rate changes; both he and Meade, by the way, confine themselves to comparative statics.

Discussions of optimal exchange rate regimes were based on the assumption that the large pre-war capital movements, characterized by political and economic speculative flights, were expected not to continue or, in any case, not be permitted. (As early as 1936 Williams outlined possible forms of international monetary and exchange arrangements.) In this light the recommendations for exchange rate flexibility of Friedman (1953), Meade (1955) and others were simple arguments in favour of permitting goods markets to clear by price variation.

The issue of the effects of internal financial policy on the foreign balance and hence on its success in achieving its domestic goals was explored with much simpler models, by Fleming (1962) and by Mundell (1960). These were initially designed to address the problem that domestic full employment policies, if successful, would worsen the trade balance. Some of their two-country models necessarily dealt with the impact on foreign countries as well. (Williams had raised this issue as early as 1934.) They produced models that dealt with the possibility of balance in external payments and attainment of a targeted level of domestic expenditure provided international financial capital flows were sufficiently elastic with respect to interest rate differentials. Metzler (1960) dealt with similar issues but concentrated primarily, in the spirit of Wicksell and Keynes, on the implications for domestic money markets and interest rates as the channel for influence on real absorption. His is a full employment model, so there is no government stabilization activity.

The question of whether monetary and fiscal authorities can maintain desired levels of inflation, real output and the real exchange rate continues to exercise the profession to the present day. Now, given the trend back into administered, if not fixed, exchange rates, the enormous stocks and flows of international financial assets, and the large current account imbalances that these permit, mirrored by the gaps between domestic savings and investment, the issue of the 'adjustment mechanism' takes the form of questions as to the sustainability of these imbalances and the consequences of diminishing or eliminating them. There is general agreement that the imbalances prevailing currently are not sustainable; the manner and consequence of their elimination is less obvious. See Clarida (2006) for an excellent summary of a National Bureau of Economic Research conference.

The overwhelming size and volatility of international financial flows (which, not by chance, coincided with the abandonment of the worldwide fixed exchange- rate system in 1973) have informed the reactivated discussions of optimum exchange rate regimes. Should rates be pegged (temporarily or in perpetuity, as in a model of dollarization) or allowed to float, freely or with some intervention? And if pegged, then to what? This leads to discussion of currency baskets, pegging to a weighted average of currencies; the question then is, what determines the weights (Flanders and Helpman 1979)? If the baskets are weighted by trade shares, international capital flows can prove highly disruptive. We are led, in turn, to the issue of whether capital movements can be controlled, and, if so, should they be, and which countries should be encouraged or permitted to attempt such controls. One line of discussion on this subject revolves around the proposed 'Tobin Tax' (Tobin 1978), designed to put a little 'sand in the wheels' of international monetary flows, which have become huge relative to commodity flows, and which can be highly erratic in response to short-run volatile shifts in expectations.

In Mundell's work the internal-external balance issue led naturally into a discussion of the requirements, in terms of both labour and capital mobility, for a group of countries to constitute an optimum currency area. (Abba Lerner had hinted at something like that in 1944.) At the time, the question was a theoretical curiosum. Twenty years later it led to substantive questions about the viability of the Eurozone as an optimum currency area, including analogies to studies of the United States as such (see Rockoff 2003), and whether a single currency and central bank can be sustained without a single fiscal authority which effects intra-area transfers.

At the same time, the recent neo-monetarism or new neoclassical trend in macroeconomics has

been paralleled by a 'new international macroeconomics'. An exhaustive and perceptive survey is provided in Lane (2001). Starting from the work of Obstfeld and Rogoff (1995), there have been numerous explorations of the impact of monetary shocks (and some consideration of technological shocks) on trade, prices, welfare, real exchange rates, real terms of trade in models of two countries, many countries, and a single small country. Some have sticky wages, all have administered prices, of various types. Different assumptions are made as to consumption elasticities, technology, non-traded goods, bias toward home goods, the inclusion of capital, financial structure and completeness of financial markets, *inter alia.* Some attempts at calibration of the models have been made, with varying success. While the spelling out of the microfoundations of international macroeconomics is intellectually satisfying, the results, as Lane, himself a contributor, avers, are 'highly sensitive to the precise denomination of price stickiness, the specification of preferences and financial market structure. For this reason, any policy recommendations emanating from this literature must be highly qualified' (Lane 2001, p. 262).

Bibliography

Arrow, K.J., H.B. Chenery, B.S. Minhas, and R.M. Solow. 1961. Capital–labor substitution and economic efficiency. *The Review of Economics and Statistics* 43: 225–250.

Bagehot, W. 1873. *Lombard Street: A description of the money market*. London: Henry S. King and Company.

Baggott, N. 1970. The modern theory of international trade and the Leontief Paradox. Ph.D. dissertation, Purdue University.

Bastable, C.F. 1889. On some applications of the theory of international trade. *Quarterly Journal of Economics* 4: 1–17.

Bastable, C.F. 1897. *The theory of international trade, with some of its applications to economic policy*, 4th ed. London: Macmillan, 1903.

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

Blecker, R.A. 1997. The 'unnatural and retrograde order': Adam Smith's theories of trade and development reconsidered. *Economica* 64: 527–537.

Brown, W.A., Jr. 1940. *The International Gold Standard Reinterpreted 1914–1934*, 2 vols. New York: NBER.

Cantillon, R. 1755. Essai sur la nature du commerce en général (written c. 1730). Ed. and trans. H. Higgs, London: Macmillan, 1931.

Caves, R.E. 1982. *Multinational enterprise and economic analysis*. Cambridge: Cambridge University Press.

Caves, R.E., and R.W. Jones. 1977. *World trade and payments*, 2nd ed. Boston: Little Brown.

Chipman, J.S. 1965–66. A survey of the theory of international trade. *Econometrica* 33, July, 477–519; 33, October, 685–760; 34, January, 18–76.

Clarida, R. 2006. G7 current account imbalances: sustainability and adjustment. Working Paper No. 12194. Cambridge, MA: NBER.

Clarida, R. (ed.). 2007. *G7 Current account imbalances: Sustainability and adjustment*. Chicago: University of Chicago Press.

Cunliffe Committee. 1918. *The interim report of the Committee on currency and exchanges after the war*. London: HMSO.

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

Dixit, A., and V. Norman. 1980. *The theory of international trade: A dual, general equilibrium approach*. Cambridge: Cambridge University Press.

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

Ellis, H.S., and L.A. Metzler (eds.). 1949. *Readings in the theory of international trade*, American Economic Association. Philadelphia: Blakiston.

Fetter, F.W. 1965. *Development of British monetary orthodoxy 1797–1875*. Cambridge, MA: Harvard University Press.

Flanders, M.J. 1974. Some problems of stabilization policy under floating exchange rates. In *Trade, stability and macroeconomics, essays in Honor of Lloyd A. Metzler*, ed. G. Horwich and P. Samuelson. New York: Academic.

Flanders, M.J. 1989. *International monetary economics 1870–1960*. Cambridge: Cambridge University Press.

Flanders, M.J., and E. Helpman. 1979. An optimal exchange rate peg in a world of general floating. *Review of Economic Studies* 46: 533–542.

Fleming, J. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9: 369–379.

Ford, A.G. 1962. *The gold standard 1880–1914: Britain and Argentina*. Oxford: Clarendon Press.

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

Friedman, M. 1953. A case for flexible exchange rates. In *Essays in positive economics*. Chicago: University of Chicago Press.

Gervaise, I. 1720. *The system or theory of the trade of the world*. London: A. Woodfall. Baltimore: Johns Hopkins University Press, 1954.

Graham, F.D. 1923. The theory of international values re-examined. *Quarterly Journal of Economics* 38: 54–86. Reprinted in Ellis and Metzler (1949).

Grossman, G.M., and E. Helpman. 1991. *Innovation and growth in the global economy*. Cambridge, MA: MIT Press.

Grossman, G.M., and E. Helpman. 2002. *Interest groups and trade policy*. Princeton: Princeton University Press.

Grubel, H.G., and P.J. Lloyd. 1975. *Intra-industry trade: The theory and measurement of international trade in differentiated products*. New York: Wiley.

Haberler, G. 1930. Die Theorie der komparativen Kosten und ihre Aüswertung für die Begründung des Freihandels. *Weltwirtschaftliches Archiv* 32: 349–370. Trans. as 'The theory of comparative costs and its use in the defense of free trade' in *Selected essays of Gottfried Haberler,* ed. and trans. A.Y.C. Koo. Cambridge, MA: MIT Press, 1985.

von Haberler, G. 1936. *The theory of international trade*. London: Hodge. English translation of German original, 1933.

Hayek, F.A. 1937. *Monetary nationalism and international stability*. New York: Augustus M. Kelley, 1964.

Heckscher, E. 1935. In *Mercantilism*, 2nd ed, ed. E.-F. Soderlund. London: George Allen & Unwin, 1955.

Heckscher, E.F. 1918, and Ohlin, B. 1924. *Heckscher–Ohlin trade theory.* Trans. and ed. with introduction by H. Flam and M.J. Flanders. Cambridge, MA: MIT, 1991.

Helpman, E., and P. Krugman. 1985. *Market structure and foreign trade*. Cambridge, MA: MIT Press.

Hollander, J.H. 1910–11. The development of the theory of money from Adam Smith to David Ricardo. *Quarterly Journal of Economics* 25: 429–470.

Hume, D. 1752. Of the balance of trade, from *political discourses*. In *David Hume: Writings on economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1955.

Jones, R.W. 2000. *Globalization and the theory of input trade*. Cambridge, MA: MIT Press.

Keynes, J.M. 1929. The German transfer problem. *Economic Journal* 39: 1–7. Reprinted in Ellis and Metzler (1949). Subsequent 'Rejoinder' and 'Reply to Ohlin and Rueff' in *Economic Journal* 39, June and September, 1929, 179–82, 404–408.

Keynes, J.M. 1937. *General theory of employment, interest, and money*. London: Macmillan.

Krugman, P. 1987. Is free trade passé? *Journal of Economic Perspectives* 1(2): 131–144.

Lane, P. 2001. The new open economy macroeconomics: a survey. *Journal of International Economics* 54: 235–266.

Leontief, W.W. 1953. Domestic production and foreign trade: The American capital position re-examined. *Proceedings of the American Philosophical Society* 97: 332–349. Reprinted in H.G. Johnson and R.E. Caves, *Readings in international trade.* Homewood: R.D. Irwin, 1968.

Lerner, A.P. 1932. The diagrammatical representation of cost conditions in international trade. *Economica* 12: 346–356. Reprinted in Lerner (1953) and Lerner (1983).

Lerner, A.P. 1933. Factor prices and international trade. Seminar paper, LSE. Reprinted in Lerner (1953) and Lerner (1983).

Lerner, A.P. 1934. The diagrammatical representation of demand conditions in international trade. *Economica* 1: 319–334. Reprinted in Lerner (1953) and Lerner (1983).

Lerner, A.P. 1944. *Economics of control*. New York: Macmillan.

Lerner, A.P. 1953. *Essays in economic analysis*. London: Macmillan.

Lerner, A.P. 1983. In *Selected economic writings of Abba P. Lerner*, ed. D.C. Colander. New York: New York University Press.

Locke, J. 1696. *Several papers relating to money, interest and trade, etcetera*, , 1968. New York: Augustus M. Kelley.

Marshall, A. 1879. In *The theory of foreign trade and other portions of economic science bearing on the principle of Laissez Faire*, ed. J.K. Whitaker. New York: The Free Press, 1975.

Marshall, A. 1923. *Money, credit and commerce*. London: Macmillan.

Martin, R., and P. Sunley. 1996. Paul Krugman's geographical economics and its implication for regional development theory: A critical assessment. *Economic Geography* 72: 259–292.

Meade, J.E. 1955. The case for variable exchange rates. *Three Banks Review* 27(September): 3–27.

Meade, J.E. 1951. *The balance of payments and the balance of payments: Mathematical supplement*. Oxford: Oxford University Press.

Metzler, L.A. 1960. The process of international adjustment under conditions of full employment: A Keynesian view. In *Collected papers*, ed. L.-A. Metzler. Cambridge, MA: Harvard University Press, 1973; and *Readings in international economics,* ed. R.E. Caves and H.G. Johnson. American Economic Association. London: Allen & Unwin, 1968.

Mill, J.S. 1848. *Principles of political economy*, , 1994. Oxford: Oxford University Press.

Mundell, R.A. 1960. The monetary dynamics of international adjustment under fixed and flexible exchange rates. *Quarterly Journal of Economics* 74: 227–257.

Mundell, R.A. 1961. A theory of optimum currency areas. *American Economic Review* 51: 657–665.

Obstfeld, M., and K. Rogoff. 1995. Exchange rate dynamics *redux*. *Journal of Political Economy* 103: 624–660.

Ohlin, B. 1924. *Handelns teori.* Reprinted in *Heckscher–Ohlin trade theory.* Trans and ed. H. Flam and M.J. Flanders. Cambridge, MA: MIT Press, 1991.

Ohlin, B. 1929. The reparation problem: A discussion, and a rejoinder to Jacques Rueff. *Economic Journal* 39. Reprinted in Ellis and Metzler (1949).

Ohlin, B. 1933. *International and interregional trade*, , 1967. Cambridge, MA: Harvard University Press.

Ricardo, D. 1811. The high price of bullion, a proof of the depreciation of bank notes. In *Pamphlets and papers 1809–1811*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1817. In *Principles of political economy and taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.
Robinson, J. 1937. Beggar-my-neighbour remedies for unemployment. In *Essays on the theory of employment*. New York: Macmillan, 1937, 1947; Reprinted in Ellis and Metzler (1949) from the 1947 edition.
Rockoff, H. 2003. How long did it take the United States to become an optimal currency area? In *Monetary unions: Theory, history, public choice: International studies in money and banking 18*, ed. F.H. Capie and G.E. Wood. London/New York: Routledge.
Samuelson, P.A. 1939. The gains from international trade. *Canadian Journal of Economics and Political Science* 5: 195–205. Reprinted in Samuelson (1966).
Samuelson, P.A. 1948. International trade and the equalisation of factor prices. *Economic Journal* 58: 163–184. Reprinted in Samuelson (1966).
Samuelson, P.A. 1949. International factor-price equalisation once again. *Economic Journal* 59: 181–197. Reprinted in Samuelson (1966).
Samuelson, P.A. 1951–2. A comment on factor-price equalisation. *Review of Economic Studies* 19: 121–122. Reprinted in Samuelson (1966).
Samuelson, P.A. 1953–4. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21: 1–20. Reprinted in Samuelson (1966).
Samuelson, P.A. 1966. In *The collected scientific papers of Paul A. Samuelson*, ed. J. Stiglitz. Cambridge, MA: MIT Press.
Sayers, R.S. 1936. *Bank of England operations 1890–1914*, , 1970. Westport: Greenwood Press.
Sayers, R.S. 1976. *The Bank of England, 1891–1944,* 3 vols. Cambridge: Cambridge University Press.
Schumpeter, J.A. 1954. *History of economic analysis.* George Allen & Unwin.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. New York: Modern Library, 1937.
Stolper, W.F., and P.A. Samuelson. 1941. *Protection and real wages*.. Reprinted in Ellis and Metzler (1949).
Stuvel, G. 1951. *The exchange stability problem*. Leyden/ New York: Augustus M. Kelley.
Taussig, F.W. 1928. *International trade*. New York: Macmillan.
Thornton, H. 1802. In *An enquiry into the nature and effects of the paper credit of Great Britain*, ed. F.-A. Hayek. New York: Augustus M. Kelley.
Tobin, J. 1978. A proposal for international monetary reform. *Eastern Economic Journal* 4: 153–159. Reprinted in J. Tobin, *Essays in economics: Theory and policy.* Cambridge, MA: MIT.
Torrens, R. 1820. *An essay on the external corn trade 1820*. London: Longman, Rees, Orme, Brown & Green.
Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper and Bros.
White, H.D. 1934. Haberler's Der internationale Handel, Ohlin's interregional and international trade. *Quarterly Journal of Economics* 48: 727–741.
Williams, J.H. 1929. The theory of international trade reconsidered. *Economic Journal* 39: 195–209. Reprinted in Williams (1944).
Williams, J.H. 1934. The world's monetary dilemma: internal versus external stability. *Proceedings of the Academy of Political Science* 16(1): 62–68. Reprinted in Williams (1944).
Williams, J.H. 1937. The adequacy of existing currency mechanisms under varying circumstances. *American Economic Review* 27(1): 151–168. Reprinted in Williams (1944) with the title 'International monetary organization and policy'.
Williams, J.H. 1944. *Postwar monetary plans and other essays*. New York: Albert A. Knopf.

International Finance

Maurice Obstfeld

I

Abstract

Fundamental to international finance is the idea of 'external balance', whereby a country's external indebtedness does not threaten its ability to meet its international obligations. The requirements of external balance have varied with the nature of the linkages among economies across historical episodes. This article both reviews the major developments in the economic analysis of external balance and traces how nations have sought to achieve it from the era of the gold standard in 19th century through the Bretton Woods system to the era of floating exchange rates that began in 1973.

Keywords

Balance of payments; Balanced trade; Bank of England; Barter trade theory; Beggar-thy-neighbour; Bretton Woods system; Central banks; Consumption smoothing; Current account; Default; Dornbusch, R.; External balance; Fiscal policy; Fisher, I.; Fixed exchange rates; Floating exchange rates; Friedman, M.; Gold standard; Gold-exchange standard; Great Depression; Hume, D.; Income–specie–flow mechanism; Inflation; Inflationary expectations; Interest rates; Internal balance;

International capital flows; International finance; International financial adjustment; International Monetary Fund; International reserves; Intertemporal maximization hypothesis; Investment; Keynes, J. M.; Liquidity preference; Lucas, R.; Meade, J.; Mercantilism; Money supply; Moral hazard; Mundell, R.; Nash–Cournot equilibrium; National income identity; Nurkse, R.; Obstfeld, M.; Ohlin, B.; Pigou, A.; Price stability; Price–specie–flow mechanism; Protectionism; Public debt; Quantity theory of money; Real-balance effect; Ricardo, D.; Samuelson, P.; Seigniorage; Sovereign debt; Sterilization; Sticky prices; Sticky wages; Terms of trade; Third World debt; Transfer problem; Triffin, R.; Unemployment; Viner, J.; Walras's law; Willingness-to-pay hypothesis

JEL Classifications
F3

International finance is concerned with the determination of real income and the allocation of consumption over time in economies linked to world markets.

Fundamental to international finance is the somewhat elusive idea of 'external balance', which in practice entails a path of external indebtedness that does not threaten a country's ability to meet its international obligations. Because the nature of the linkages among economies has varied across historical episodes, the requirements of external balance have varied as well. International finance studies the policies and market forces which may lead to external balance under various conditions. The history of the subject illustrates how the nature of world market linkages has itself been changed by national efforts to cope with external constraints.

The national income identity is the necessary groundwork for any discussion of external balance. The national income of an open economy equals domestic product plus net factor payments from abroad plus net international transfer payments; the current account equals net exports of goods and services (including all net factor payments) plus net transfers. If national expenditure is defined as the sum of consumption and investment (by both the public and private sectors), the national income identity asserts that national income less national expenditure equals the current account. When in surplus, the current account therefore measures the growth of the economy's external assets; when in deficit, it measures the growth of external debt.

The Classical Paradigm

The classical Ricardo–Mill barter trade theory shows how the terms of trade and international production pattern are determined in a stationary world economy with balanced trade. The classical analysis of the transition to balanced trade may be viewed as an account of the convergence process to the long-run barter equilibrium. As Ricardo noted in the *Principles* (1817):

> Gold and silver having been chosen for the general medium of circulation, they are, by the competition of commerce, distributed in such proportions amongst the different countries of the world as to accommodate themselves to the natural traffic which would take place if no such metals existed, and the trade between countries were purely a trade of barter.

Historically, however, the classical paradigm of external adjustment preceded Ricardo. Major elements of the theory had been expounded quite clearly by the early 18th century, but the most coherent and effective exposition was given by Hume in 1752.

Hume assumed a world economy that settles trade imbalances exclusively through imports or exports of precious metals that also serve as money. Building on the quantity theory of money, he constructed a full dynamic model of the balance of payments and the terms of trade. The famous price–specie–flow mechanism was put forth as an automatic market process that always works to restore balanced trade.

Hume's goal was to refute mercantilist and protectionist arguments by showing that market forces would ensure in the long run a 'natural' distribution of specie among countries.

Hume invited his readers to imagine that four-fifths of Great Britain's money supply were 'annihilated in one night'. British prices would naturally fall, he argued, cheapening British exportables relative to foreign goods and creating a trade surplus. As a result of this surplus Britain would accumulate foreign wealth in the form of specie, seeing its money supply, and hence its prices, rise. Abroad, the drain of specie would lower prices. Britain's trade surplus would dwindle and eventually disappear once its terms of trade had improved sufficiently, and at this point, the natural distribution of specie would prevail. A hypothetical fivefold increase in Britain's money supply would set off the reverse process, involving an initial improvement in Britain's terms of trade and a trade balance deficit. Over time, specie would flow abroad as the terms of trade deteriorated and external equilibrium was restored.

There is little exaggeration in saying that issues raised by Hume's analysis dominated writing in international finance up until the inter-World War years. In a period that culminated in the classical gold standard, it was natural to take as the benchmark of external balance an absence of international specie movements. Hume had placed relative price movements at the centre of his account of how external balance would be attained, but subsequent writers asked whether direct income or wealth effects might also be operative, and whether external adjustment could take place in some cases without price changes. Such questions arose in the 1929 Keynes – Ohlin debate over the German transfer problem, but as Viner (1937) showed, the questions had been raised much earlier.

A simple model of a Humean world makes apparent some of the assumptions underlying the price–specie–flow mechanism. Such a model also serves as a springboard for understanding later developments in the analysis of external adjustment. (A more detailed exposition of a similar model is given by Dornbusch (1973), whose analytical approach is, however, somewhat different from that taken here).

Assume a world of two countries, each specialized in the production of a single commodity that is consumed in both countries. With given supplies of capital and labour within each country and perfect wage flexibility, home-country output is fixed at the full-employment level x and foreign-country output is fixed at y. Let q denote the price of y-goods in terms of x-goods (the terms of trade), z domestic expenditure measured in x-goods, and z^* foreign expenditure, also measured in x-goods. Then the domestic demands for the two goods are $c_x(q, z)$ and $c_y(q, z)$, while the foreign demands are $c_x^*(q, z^*)$ and $c_y^*(q, z^*)$.

Expenditure is determined by monetary conditions. The money supplies M and M^* are for simplicity taken to consist entirely of gold, and P and P^*denote the money prices of home and foreign goods, respectively. The exchange rate between domestic and foreign currency can be set at unity with no loss of generality, so the terms of trade, q, equal P^*/P. In each country there is a desired long-run (or 'natural') money supply: this is proportional to nominal output, and saving behaviour is governed by discrepancies between natural and actual money supplies. Because a country's net saving here equals its current account, which by assumption is settled in specie, saving behaviour determines the evolution of national money supplies.

These evolve according to the laws

$$\begin{aligned} \mathrm{d}M/\mathrm{d}t &= \theta(\chi Px - M)\mathrm{d}M^*/\mathrm{d}i \\ &= \theta^*(\chi^* P^* y - M^*), \end{aligned}$$

where $\chi(\chi^*)$ is the reciprocal of the home (foreign) country's long-run monetary velocity and $\theta(\theta^*)$ is the home (foreign) marginal propensity to dissave out of monetary wealth. Expenditure levels are therefore

$$\begin{aligned} z &= (1 - \theta\chi)x + \theta M/P, \quad z^* \\ &= (1 - \theta^*\chi^*)qy + \theta^* M^* P, \end{aligned}$$

where $\theta\chi$, $\theta^*\chi^* < 1$.

The model is closed by two equilibrium conditions. With a given world stock of monetary gold, M^w, home saving must equal foreign dissaving, that is, world expenditure must equal world output. In addition, the market for domestic goods must clear. By Walras's Law, these two

equilibrium conditions imply equilibrium in the market for foreign goods.

The condition of zero desired world saving is $(dM/dt) + (dM^*/dt) = 0$, or

$$P = \frac{\theta M + \theta^*(M^w - M)}{\theta\chi x + \theta^*\chi^* qy} \quad (1)$$

Equation 1 shows that, for given terms of trade and money supplies, the world price level adjusts to maintain consistency between the countries' saving plans. In equilibrium, this condition makes P a function of q and $M, P = P(q; M)$, with

$$\frac{q}{P}\frac{\partial P}{\partial q} = \frac{-\theta^*\chi^* qy}{\theta x^x + \theta^*\chi^* qy} > -1\,\frac{M}{P}\frac{\partial P}{\partial M} = \frac{(\theta - \theta^*)M}{\theta xPx + \theta^*\chi^* P^* y} \lesseqgtr 0.$$

The market for x-goods clears when

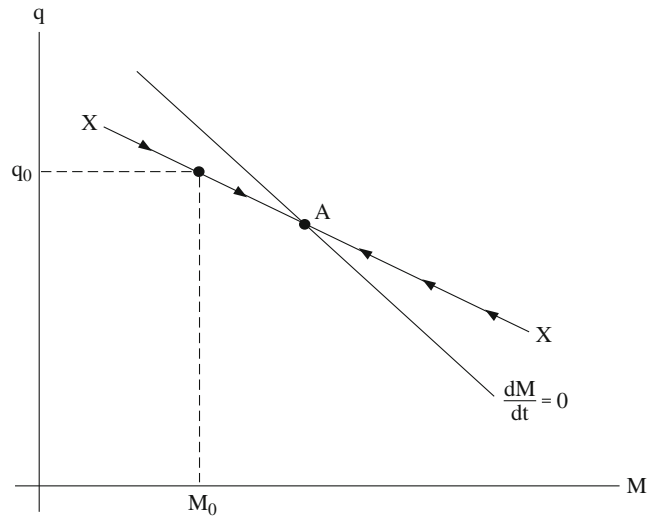
$$c_x[q, (1 - \theta\chi)x + \theta M/P] + c_x^*[q, (1 - \theta^*\chi^*)qy + \theta^* M^*/P] = x \quad (2)$$

Substitution of $P = P(q, M)$ and $M^* = M^w - M$ into (2) gives the curve describing combinations of M and q at which both goods markets clear and aggregate world saving is zero. The curve is labelled XX in Fig. 1 and is shown with a negative slope. The assumptions giving rise to this negative slope are crucial for analysing the Humean adjustment process. An increase in M (which necessarily implies an equal fall in M^*) causes an excess demand for x-goods equal to

$$\frac{\theta\theta^*(M + M^*)}{P(\theta M + \theta^* M^*)}\left[\partial c/\partial z - \partial c_x^*/\partial z^*\right]$$

near the system's long-run equilibrium (where $dM/dt = dM^*/dt = 0$). The term is the $\partial c_x/\partial z - \partial c_x^*/\partial z^*$ difference between the two countries' marginal propensities to spend on home-country goods; if the home-country marginal propensity is larger – the 'orthodox' presumption in transfer analysis (Samuelson 1971) – a redistribution of nominal balances in favour of the home country creates an incipient excess demand for its output. This excess demand is eliminated by a fall in q if the home- goods market is Walras stable, so XX slopes downward under standard assumptions concerning marginal spending propensities and Walrasian stability.

The curve in Fig. 1 labelled $dM|dt = 0$ describes points at which $M = {}_{\chi}p(q, M)x$. This locus has a negative slope algebraically smaller than that of XX. With goods markets continuously in equilibrium, the world economy travels along XX to its long-run equilibrium at point A, where



International Finance, Fig. 1

international prices and the distribution of specie give rise to balanced trade.

In most respects the model confirms Hume's account of the external adjustment process. A (small) fall in M to M_0, for example, leads to terms of trade q_0, which are worse for the home country. The terms-of-trade change is a direct result of the transfer of purchasing power to foreigners, which produces an excess supply of home goods at the initial prices. The home balance-of-payments surplus that simultaneously emerges causes a gradual redistribution of money in favour of the home country, so the home terms of trade improve during the transition to external balance.

If $\theta = \theta^*$, equilibrium P is a function of q alone, with a negative elasticity greater than -1. The rise in q caused by a fall in M is thus accompanied by a less- than-proportional fall in P and a rise in P^* that are reversed as the economy returns to point A. These results are in accord with Hume's predictions, but they need not hold if the expenditure responses to real balances differ sufficiently in the two countries. If $\theta > \theta^*$, a transfer of money abroad raises world saving for given terms of trade, so P^* may *fall* along with P and then rise during the subsequent adjustment. Likewise, if $\theta < \theta^*$, a money transfer abroad may reduce world saving sufficiently that P must *rise*, along with P^*, to restore goods-market equilibrium in the short run. In this case, the initial response to the disturbance is followed by price deflation in both countries.

This stylized version of Hume's paradigm may be used to analyse the transfer problem. Suppose that ownership of a portion of the foreign country's endowment is given to the home country. Does the home trade deficit necessarily increase by the amount of the transfer, or is the transfer undereffected, requiring a flow of specie to the home country to balance international accounts? A second focus of debate in the literature is the possibility that the transfer imposes a 'secondary burden' on the paying country by adding an equilibrium terms-of-trade deterioration to the primary income burden. Keynes and Ohlin clashed on this point in 1929, with Keynes arguing that the secondary burden is inevitable.

To simplify, suppose that $\theta = \theta^*$ and $\chi = \chi^*$. Since long-run money demand rises with income, the $dM/dt = 0$ locus shifts to the right, implying that the transfer at first is undereffected and that the world's gold stock is redistributed toward the home country. Under the standard assumption regarding marginal spending propensities, the transfer also creates an excess demand for x-goods at the initial terms of trade, so XX shifts downward. A secondary burden is thus imposed on the paying country, and this burden worsens over time as balanced trade is reestablished.

The Interwar Period

The years between the World Wars saw a partial and ultimately unsuccessful return to the gold standard, followed by extensive experimentation with floating exchange rates and direct controls on international payments as means of attaining external balance. Nurkse's (1944) account of the period is probably the most influential one. Writers on international finance continued to conceptualize external balance in terms of reserve movements. The spread of the gold-exchange standard, under which central banks held as foreign reserves currencies tied to gold as well as gold itself, broadened the class of assets through which balance-of-payments deficits were financed.

International capital movements were discussed increasingly in the theoretical literature, but they were viewed for the most part as an adjunct to the classical balance-of-payments adjustment mechanism. The theoretical discussions merely formalized a mechanism that had long been exploited by the Bank of England to regulate gold flows. A country that suddenly developed a trade deficit would face declining international reserves, a declining money supply, and rising interest rates. Rising interest rates would, however, attract foreign capital inflows and thus dampen the resulting deficit in the balance of payments. On this view, interest-sensitive capital flows had a potentially stabilizing role to play in discouraging protracted reserve flows. Given the turbulent conditions of the period, contemporary writers fully recognized that capital

flows motivated by fears of devaluation or political instability could just as well destabilize an already bad external payments problem.

Such 'short-term' or interest-sensitive capital movements were generally discussed separately from 'long-term' international capital movements which directly financed investment or government expenditures. Theoretical discussions of long- term capital movements focused mainly on the transfer mechanism, the balance-of- payments and terms-of-trade adjustments that would accompany an inter-country transfer of capital. Conspicuously absent from the literature were attempts to develop a normative intertemporal theory of international capital transfer. Such a theory naturally would have extended the prevailing external balance concept to comprise changes in nations' *overall* indebtedness rather than just changes in the central bank's foreign assets. It had been known, at least since Ricardo's *Principles*, that producers and consumers could gain if long-term foreign investment equalized profits internationally. The insight did not dominate thinking about the nature of external balance.

This gap in the literature is surprising in view of the developments in international capital markets over the previous century. Huge flows of long-term capital, primarily from Britain, had financed railroad construction and other investment in the Western Hemisphere. France and Germany also made significant foreign loans. In the early 1930s, widespread foreign debt default among the Latin American countries highlighted the need to analyse formality the sustainability of external debt paths. In the world assumed by Hume, specie flows had been the only means of settling current-account imbalances, and a concept of external balance based on balance-of-payments equilibrium had been defensible. Such a concept of external balance was outmoded, however, in a world where other types of asset trade could finance the current account.

The necessary change of perspective did not occur for several decades. Instead, the events and ideas of the interwar period led international financial theory to turn away sharply from the concern with the dynamics of international adjustment underlying the classical model. Emphasis shifted inward, to the interaction between the balance of payments and domestic economic conditions.

The Bretton Woods Period

The interwar experience had a profound influence on both the institutional framework of postwar international finance and the theoretical orientation of researchers. The international agreement reached at Bretton Woods in 1944 set up a world trading community linked by fixed dollar exchange rates, with a United States commitment to peg the dollar price of gold at $35 per ounce providing an anchor for the world price level. The agreement's provisions aimed to promote free trade in goods, but private capital movements were viewed as potentially disruptive and the widespread capital controls then in force were not discouraged. A prevailing view that flexible exchange rates had failed during the interwar period motivated the adoption of a fixed-rate system. Provision was made, however, for infrequent exchange-rate adjustment, after due consultation, in circumstances of 'fundamental disequilibrium' in the balance of payments.

Central to the design of the Bretton Woods system was a desire to avoid unemployment and ensure price-level stability. In the interwar years, many governments had resorted to competitive currency depreciations and trade restrictions aimed at reducing domestic unemployment. These 'beggar-thy-neighbour' moves made all countries worse off. Having recently experienced the hardships of the worldwide Great Depression, the Bretton Woods signatories recognized the goal of 'internal balance' – full employment with price stability – as a key aim of government policy. An International Monetary Fund was set up to reconcile the goals of internal and external balance. It was hoped that the availability of Fund credit would make it unnecessary for members to tolerate high unemployment in pursuing external balance, or to interfere with trade flows in pursuing internal balance.

In an environment of fixed exchange rates and extremely limited capital mobility, the overriding

external consideration for governments was the available stock of foreign, particularly dollar, reserves. The operative external target was therefore the acquisition of as many dollars as possible through balance-of-payments surpluses. As the reserve centre, the United States enjoyed the privilege of being able to finance its own balance-of-payments deficits by borrowing dollars from foreign central banks. In reality, however, the United States was not totally free of a reserve constraint. Foreign central banks could, and did, use their dollars to buy gold from the US authorities at the official price. The problem of gold losses became important as the postwar period of 'dollar shortage' ended in the late 1950s. In 1960, Triffin put the American external dilemma in its most sombre light: Once foreign official dollar holdings exceed the official value of the US gold stock, it would become impossible to satisfy all foreign claims to US gold without a rise in the dollar price of metal. The resulting confidence problem, Triffin predicted, would undermine the stability of the Bretton Woods system.

As it developed immediately after World War II, international financial theory reflected the new institutional arrangements, along with the economic assumptions underlying Keynes's (1936) diagnosis of the unemployment of the 1930s. The new paradigm, set forth very effectively by Metzler (1948, pp. 212–13), assumed sticky price levels and wages along with fixed exchange rates, thus precluding the relative- price adjustments at the heart of the classical paradigm while opening the door to employment fluctuations:

> The important feature of the classical mechanism ... is the central role which it attributes to the monetary system. The classical theory contains an explicit acceptance of the Quantity Theory of Money as well as an implied assumption that output and employment are unaffected by international monetary disturbances. In other words, the classical doctrine assumes that an increase or decrease in the quantity of money leads to an increase or decrease in the aggregate money demand for goods and services, and that a change in money demand affects prices and costs rather than output and employment ... . The essence of the new theory is that an external event which increases a country's exports will also increase imports *even without price changes*, since the change in exports affects the level of output and hence the demand for all goods. In other words, movements of output and employment play much the same role in the new doctrine that price movements played in the old.

An increase in external demand for a country's exports, for example, would raise the country's trade surplus in the first instance, but once the multiplier effect of the disturbance had raised income and hence import spending, the initial impact on the trade balance would be reduced. Metzler noted, however, that even if one assumed that investment spending responds positively to a rise in real income, it was unlikely that multiplier effects alone would ensure complete trade-balance adjustment in the short run.

The Keynesian account of external adjustment therefore contained an important gap. Private capital movements were largely ruled out in the Keynesian models, so incomplete trade-balance adjustment implied incomplete balance-of-payments adjustment and growing or shrinking central-bank foreign reserves. The models pushed monetary factors to the background, implicitly or explicitly assuming that central-bank sterilization operations were offsetting any monetary effects of the balance of payments. Only a few of the early postwar theorists, notably Meade (1951), assigned an important role to monetary factors.

Even if the sterilization assumption were granted, however, consideration of the system's inherent dynamics made clear the infeasibility of a *permanent* sterilization policy. Countries with persistent deficits would ultimately exhaust their available international reserves, including IMF credit; and even surplus countries might be unable to sterilize indefinitely if domestic financial markets were thin. How, then, could trade-balance equilibrium even be restored after a permanent external shock? Fiscal policy could be effective in situations where the needs of internal and external balance were both served by the same measure. In dilemma situations where fiscal measures could move the economy toward external balance only at the cost of increasing its distance from internal balance, the 'fundamental disequilibrium' clause of the IMF Articles of Agreement could be invoked and the currency devalued. But no

automatic market mechanism pushing the economy toward balance-of-payments equilibrium was featured in the early postwar writing.

In a series of remarkable papers published in the early 1960s, Mundell revived the explicit dynamic analysis of international adjustment. His models placed the monetary sector in the foreground, adopting a Keynesian liquidity-preference view of interest-rate determination. A prescient paper by Metzler (1960), written at about the same time, took a similar approach.

Mundell's paper on 'The International Disequilibrium System' (1961) criticized the Keynesian model's failure to account for the dynamic effects of payments imbalances. Even in a Keynesian world, Mundell argued, an income–specie–flow mechanism, analogous to Hume's price–specie–flow mechanism, ensures long-run balance-of-payments equilibrium. A 'fivefold increase' in a country's money supply, for example, depresses domestic interest rates, stimulates investment spending, and creates a deficit in the balance of payments. As the central bank loses reserves, however, the interest rate gradually rises and reduces investment, the process coming to an end (for a small country) only when the domestic money supply, the interest rate, investment, and output have returned to their original levels. The introduction of dynamic adjustment made it clear that sterilization could have only limited success as a policy response to permanent balance-of-payments disturbances. One source of dynamic effects, however, was not explicitly analysed in Mundell's work of the period. The omitted effect was the real-balance effect on expenditure, central to the classical account but possibly relevant (as Pigou had shown) under Keynesian conditions as well.

In line with the increasing international capital mobility that followed the European move toward currency convertibility in 1958, Mundell gave the capital account a prominent role in his models. The presence of capital mobility suggested a solution to the policy dilemmas that could arise under fixed exchange rates when the goals of internal and external balance appeared to conflict. Mundell showed that by gearing monetary policy to external balance and fiscal policy to internal balance, governments could simultaneously attain both goals. The key to the argument is the observation that monetary and fiscal expansion both raise output but have different effects on the capital account, monetary expansion causing capital outflows (by driving down the home interest rate) and fiscal expansion causing capital inflows (by raising the interest rate). With two independent instruments, both internal and external policy targets can be attained simultaneously.

While a major step forward, the Mundellian argument for a policy mix suffered from two drawbacks. First, the theoretical specification of the capital account as a function of international interest-rate levels was weak: it seemed unlikely that capital would flow at a uniform level forever even if the interest differential remained fixed. Missing was a discussion of stock equilibrium in international asset markets. The second problem with the policy mix was its definition of external balance. Would any policymaker view with satisfaction a permanently high interest rate that brought about balance-of-payments equilibrium by crowding out domestic investment and encouraging a build-up of external debt? Key considerations omitted from Mundell's model were the stock of net foreign claims and the associated flows of interest payments. Mundell himself (1968, p. 207) recognized that in many contexts, the definition of external balance as balance-of-payments equilibrium might be inadequate:

> Just as the composition of output is important (the division of output between investment and consumption affects additional growth targets), so an appropriate composition of the balance of payments is a legitimate target of policy.

Indeed, in spite of the continuing obligation to peg dollar exchange rates, the standard definition of external balance was becoming increasingly outmoded by the late 1960s. The balance of payments remained a legitimate concern, of course, in part because a large or persistent imbalance might look like 'fundamental disequilibrium' to the market and spark a speculative attack on the currency involved. But the increasing integration of national financial markets – a development epitomized by the growth of Eurocurrency trading – weakened

the bite of the balance-of-payments constraint. In a hypothetical world of *perfect* capital mobility, a central bank short on reserves can essentially borrow them from abroad at no net cost simply by contracting domestic credit. Such an action, by causing an incipient rise in the home interest rate, leads to an instantaneous private capital inflow and an official reserve gain equal to the fall in domestic credit. The home interest rate, the money supply, output, and the national external debt are unchanged in the final equilibrium: the central bank holds more foreign assets and fewer domestic assets, while the home private sector, having made the mirror-image adjustment, holds fewer foreign assets and more domestic assets.

The case of perfect capital mobility is an extreme one that does not fit the facts of the late Bretton Woods period. None the less, the opportunities for central banks to borrow dollar reserves in the international capital market had grown since the early 1960s. The situation facing the United States was quite different. As the primary international reserve issuer, its responsibility was to peg the dollar price of gold, a responsibility that would have required the gearing of US monetary policy to that external commitment. In spite of such expedients as the two-tier gold market established by central banks in 1968, the US did not succeed in preserving the dollar's link to gold. Triffin had been right. After a series of violent speculative attacks, the US severed the dollar's gold link in August 1971 and in December 1971 devalued the dollar against major foreign currencies. The patchwork system of fixed exchange rates proved unstable, and in the first months of 1973 the postwar period of floating exchange rates began.

Floating Exchange Rates

The industrialized countries adopted floating dollar exchange rates as an interim measure, but in fact a significant body of economists had come to advocate floating rates by 1973. Friedman's (1953) powerful case for flexible rates was the opening shot in a campaign to revise the then-prevailing view, expounded by Nurkse (1944), that the floating-rate experiments of the interwar years were disastrous. By the time Johnson wrote his well-known polemic of 1969, Friedman's views had gained many adherents.

The fundamental argument for floating rates was that they would free governments of the balance-of-payments constraint and allow them to use monetary policy to attain domestic economic goals. Equilibrium in the balance of payments would be automatic if central banks simply refrained from intervening in the foreign exchange market. At the same time, floating rates would permit central banks to target their nominal money supplies without being frustrated by offsetting interest- sensitive foreign reserve flows. Widespread restrictions on trade and capital movements, motivated in part by a desire to impede reserve flows under the fixed- rate regime, could be dismantled.

Subsequent experience was to provide only partial vindication to the advocates of floating. In the decade after 1973, barriers to capital movement were reduced to insignificant levels in many of the industrial countries. This development helped spark unprecedented growth in international financial intermediation. Under the new exchange-rate regime, however, policymakers became more acutely aware that the traditional definition of internal balance as full employment cum price stability really involved two, quite distinct, goals. Under a floating exchange rate, monetary expansion aimed at domestic unemployment translates immediately into currency depreciation, higher import prices, and heightened inflationary expectations.

Conversely, a rapidly adjusting exchange rate provides a powerful channel through which inflationary expectations can have a direct and immediate effect on inflation in an open economy. Any short-run tradeoff between inflation and unemployment would therefore be less favourable under a floating rate. Floating rates certainly allow countries to choose their own trend inflation rates. But it soon became evident that if disturbances to the economy originated predominantly outside the money market, the inflationary cost of using monetary policy to target employment could be quite high.

Sharp exchange-rate movements might also have adverse distributional effects in the economy, and these, together with a desire for price-level stability, led central banks to intervene, at times heavily, in the foreign exchange market.

Correspondingly, the predicted drop in central banks' demand for international reserves did not materialize (although the composition of reserves did change over time as the Deutschmark and yen became important reserve currencies and the pound sterling retreated). Central banks' use of foreign reserves to manage exchange rates did not necessarily imply an operative balance-of-payments constraint, however, since in many countries the same exchange-rate effects could have been achieved at an unchanged reserve level through domestic credit measures.

Under conditions of limited capital mobility, such as those existing in the early 1950s when Friedman wrote, the automatic balancing of international reserve movements by a floating exchange rate amounted essentially to the automatic balancing of the current account. With means other than reserve flows available to settle current-account imbalances, however, there is no theoretical necessity for a floating rate to balance the current account in the short run. A current-account deficit, say, can be financed entirely through domestic borrowing abroad with no decline in the central bank's foreign assets. Experience was to show that floating exchange rates themselves could not prevent the emergence of large and persistent current-account imbalances. These imbalances were problematic not only because they usually entailed costs of shifting productive resources between the economy's tradable and nontradable sectors, but also because they implied changes in foreign debt and thus in sustainable future consumption levels.

Attention therefore shifted to the mechanism of current-account adjustment under floating exchange rates and capital mobility, with researchers asking, as Hume had, if market forces would automatically push economies toward current-account balance. The new generation of dynamic open-economy models produced in the mid-1970s built on a number of antecedents in the literature. One of these was the neoclassical monetary approach to the balance of payments, which stressed the real balance effect and the transition to long-run payments equilibrium (see, for example, Frenkel and Johnson 1976). The second important antecedent was the closed- economy literature on money and growth, which had clarified the stock-flow distinction in multi-asset models with wealth accumulation. As suggested by the rational-expectations revolution in macroeconomics, many model builders endowed agents with forward-looking exchange-rate expectations that played a key role in clearing the asset markets.

The intrinsic dynamic mechanism in these models is fuelled by wealth, broadly defined to include not only real monetary balances, but also foreign assets and possibly capital, physical as well as human. (See Obstfeld and Stockman 1985, for a survey.) In line with the long-run nature of the inquiry, the 'classical' conditions of price flexibility and full employment were generally assumed, giving a productions structure similar to the Humean model set out above. Where the models differed essentially from Hume was in the wider spectrum of marketable assets, and in the resulting portfolio problem of private agents. Each given configuration of world asset stocks determines a short-run equilibrium defined by the requirements of market clearing in asset as well as goods markets. The resulting equilibrium wealth levels and real interest rates determine consumption levels at home and abroad, but there is no necessary requirement of current-account balance in the short run: goods-market equilibrium implies only that one country's planned current-account surplus equals the other's planned current-account deficit. The international adjustment process can now be visualized. All else equal, the deficit country is running down its wealth by borrowing from abroad, so its consumption is falling and foreign consumption is rising. Under the orthodox transfer criterion, this redistribution of wealth between the countries causes the deficit country's terms of trade to deteriorate over time; if anticipated, the evolution of the terms of trade has further repercussions on world real interest rates and expenditure levels. The process comes to an end once the deficit country's consumption has fallen into line with its income, which is lower than initially because of the increased interest

burden of the external debt. (A very similar adjustment process would take place with mobile capital and a fixed exchange rate, but reserve movements rather than exchange-rate movements would contribute to asset-market balance during the transition to long-run equilibrium.)

This simple picture of the adjustment process becomes more complicated once domestic capital accumulation is allowed. A current-account deficit may now finance an investment boom in which the deficit country's terms of trade improve over time. Eventually, however, the international wealth-flow mechanism restores a balanced current account. Further complications arise when the classical assumptions are dropped and Keynesian price stickiness in output markets is assumed. In such models, the approach to the long-run, full-employment equilibrium can be oscillatory.

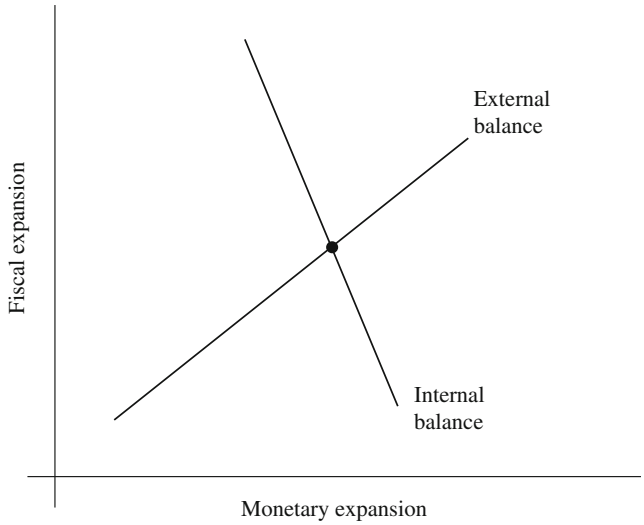
For a single economy with Keynesian features, there is an analogue to the Mundellian idea of using monetary and fiscal policy simultaneously to attain internal and external targets. Figure 2, which is developed more fully in Obstfeld (1985, pp. 408–10), illustrates this approach. The downward sloping internal-balance schedule shows combinations of monetary and fiscal ease consistent with full employment. On the assumption that monetary ease improves the current account by depreciating the currency, the external-balance schedule, which shows policy settings consistent with some current-account target, slopes upward. The intersection of the two schedules shows how policies should be set to achieve both of the government's goals in the short run.

Even if one leaves aside the complex game-theoretic problems surrounding interactions between expectations and policy, the usefulness of the above framework as a normative guide is limited by its failure to incorporate some key dynamic elements. If the government can hit its targets only by running a budget deficit, its fiscal stance must eventually be reversed if the government debt is to be serviced. In addition, the policy equilibrium shown in Fig. 2 may imply a domestic investment rate that is socially sub-optimal. Finally, the framework itself gives no guidance as to the appropriate external-balance criterion. The balanced current account reached in the hypothetical long-run equilibrium of a stationary world economy may be far off the mark in the short run in which policy decisions must be made. Recently, the theory of international finance has made partial progress in addressing these issues.

The Intertemporal Analysis of External Balance

In the 1980s, it became increasingly common to analyse the dynamic behaviour of open

International Finance, Fig. 2



economies in terms of the intertemporal maximization hypothesis applied by Fisher (1930) to the theory of saving and investment. As usual, this trend was the result of both new theoretical approaches in macroeconomics generally and of economic events that existing open-economy models seemed ill-equipped to analyse.

Lucas's (1976) influential critique of econometric policy evaluation was important in motivating the intertemporal approach. Lucas argued that the standard econometric models of the time would generally not be invariant to policy changes. Because the parameters estimated were not the 'deep' parameters describing preferences or technology, but instead reflected both deep structure and the policy environment prevailing over the estimation period, the models could not be used to analyse *changes* in the policy environment. Lucas's analysis suggested that more reliable policy conclusions might be drawn from open-economy models if demand and supply functions were derived from the optimal decision rules of maximizing households and firms.

Further impetus to develop an intertemporal approach came from events in the world capital market, particularly the international pattern of current accounts following the sharp oil-price increases of 1973–1974 and 1979–1980. The divergent patterns of a current-account adjustment by industrialized and developing countries raised the inherently intertemporal problem of characterizing the optimal response to external shocks. Neither classical nor Keynesian transfer analysis offered any reliable guidance on this question. Similarly, the explosion in bank lending to developing countries after the first oil shock sparked fears that some countries' external debt burdens would become unsustainable. The need to assess developing-country debt levels again led naturally to the notion of an intertemporally optimal current-account deficit.

Any intertemporal analysis of external balance must begin by specifying the economy's technological and market opportunities for shifting consumption over time. These opportunities are described by the economy's intertemporal budget constraint, which specifies the terms on which the economy can borrow or lend abroad, as well as the domestic investment technology. Separate analysis of the public and private sector's budget constraints illuminates the link between the public finances and external imbalances, as measured by the balance of payments or by the current account. The economy-wide budget constraint results from consolidation of the public- and private-sector constraints.

Assume for simplicity that a single good is consumed and produced on each date, and consider the position of a small open economy that can borrow or lend internationally at the real interest rate ρ. For each date t, the government of the economy chooses a level of real government consumption, $g(t)$, and a (possibly negative) level of real transfers to the private sector, $\tau(t)$. The government finances its outlays by issuing debt, by printing money, and by drawing on the interest paid by the central bank's foreign reserves. (For present purposes, the central bank's budget is best viewed as a component of the government's budget). Let $b^{G}(t)$ denote real government bond holdings (other than central-bank foreign reserves), $D(t)$ the money value of central-bank domestic credit, $P(t)$ the money price level, and $r(t)$ real foreign reserves. If the government pays the interest rate ρ on the public debt $[-b^{G}(t)]$, then the path of government bond holdings satisfies the equation:

$$\begin{aligned} \mathrm{d}b^{G}(t)/\mathrm{d}t = {} & \rho\left[b^{G}(t) + r(t)\right] \\ & + [1/P(t)]\mathrm{d}D(t)\mathrm{d}t - g(t) \\ & - \tau(t). \end{aligned} \tag{3}$$

Changes in the economy's money supply, $M^{s}(t)$, result from changes in the central bank's foreign or domestic assets. If the world price level P^{*} is constant (so that proportional changes in $P(t)$ equal proportional changes in the exchange rate), then the central-bank balance-sheet identity implies $\mathrm{d}M^{s}(t)/\mathrm{d}t = p(t)[\mathrm{d}r(t)/\mathrm{d}t] + \mathrm{d}D(t)/\mathrm{d}t$. Let $m(t)$ denote the private sector's desired real money balances and $\pi(t)$ the home inflation rate. On the assumption that the money market is continuously in equilibrium, $m(t) = M^{s}(t)/p(t)$ and Eq. 3 becomes

$$\mathrm{d}\left[b^{\mathrm{G}}(t)+r(t)\right]/\mathrm{d}t=\rho\left[b^{\mathrm{G}}(t)+r(t)\right]+\pi(t)m(t)+\left[\mathrm{d}m(t)/\mathrm{d}t\right]-g(t)-\tau(t). \quad (4)$$

Integrate (4) forward from $t = 0$ and impose the condition $\lim_{t\to\infty}\exp(-\rho t)[b^{G}(t) + r(t) \geq 0$, which restricts the government to borrowing paths such that the public debt is assymptotically paid off. The result is the intertemporal budget constraint of the government,

$$\int_0^\infty [g(t)+\tau(t)]\exp(-\rho t)dt \leq \int_0^\infty [\pi(t)m(t)+dm(t)/dt]\exp(-\rho t)dt + b^{G}(0)+r(0).$$

The inequality states that the present value of net government outlays must be less than the present value of the seigniorage from money creation plus the government's initial asset position. The latter quantity, in turn, equals central-bank foreign reserves less the public debt. For a world of perfect capital mobility, the constraint makes clear that it is the government's overall asset position that is relevant for assessing solvency. The level of foreign reserves $r(0)$ has little significance in itself. As noted earlier, the central bank can increase its reserves by selling other government assets (thus reducing $b^{\mathrm{G}}(0)$ by an amount equal to the rise in reserves). The transaction requires no change in the path of planned government outlays, $g(t) + \tau(t)$.

Consider next the private sector. Let $b(t)$ denote net private real bond holdings and $k(t)$ real capital holdings. (By assumption capital's real price equals unity). Foreigners do not hold domestic money or capital, although the analysis could easily be modified to account for these possibilities. Given an inelastic labour supply normalized at unity and a neoclassical production function $x[k(t), t]$, private-sector assets obey the equation

$$\mathrm{d}[\mathrm{b}(t)+k(t)+m(t)]/\mathrm{d}t = x[k(t),t]+\rho b(t)+\tau(t)-c(t)-\pi(t)m(t). \quad (5)$$

Define investment $i(t)$ as $\mathrm{d}k(t)/\mathrm{d}t$. The sum of (4) and (5) is

$$d\left[b(t)+b^{\mathrm{G}}(t)+r(t)\right]/dt = x[k(t),t]+\rho\left[b(t)+b^{\mathrm{G}}(t)+r(t)\right]-c(t)-i(t)-g(t).$$

The sum $b(t) + b^{\mathrm{G}}(t) + r(t)$ will be denoted by $f(t)$: $f(t)$ equals the economy's overall net claims on the rest of the world. Integrated forward and combined with the condition $\lim_{t\to\infty}\exp(-\rho t)f(t) \geq 0$ the above equation implies the economy's overall intertemporal budget constraint,

$$\int_0^\infty \{c(t)+i(t)+g(t)-x\left[k(t),t\right)\}\exp(-\rho t)\mathrm{d}t \leq f(0). \quad (6)$$

(The same constraint is relevant when the private sector is prohibited from transacting in the world capital market, but the paths of consumption, investment, and output would generally change if such a prohibition were imposed.)

Inequality (6) states that the present value of the economy's expenditures cannot exceed the present value of output plus initial net external assets. Alternatively, (6) constrains the present value of the economy's trade balance deficits to its initial foreign asset stock. The initial foreign asset stock thus limits the economy's ability to maintain absorption levels in excess of output.

An implication of the analysis is that the most appropriate indicator of flow disequilibrium in external transactions is the change in the economy's overall external assets – the current account. A surplus in the balance of payments may indicate low domestic credit expansion or growing domestic money demand; but when the government has unlimited access to the world capital market, a growing stock of foreign reserves is, in itself, neither a necessary nor a sufficient condition for a sound external position.

The important consequences of current-account flows do not imply that external balance and current-account balance are the same. In analogy with the idea of a high-employment government budget surplus, external balance could be defined roughly as a current account that

maintains the highest possible steady consumption level consistent with the economy's expected intertemporal budget constraint. (A more exact definition would require a more explicit treatment of the preferences of households and the government). Temporary unfavourable movements in output, world interest rates, or the terms of trade are appropriately offset by temporary current-account deficits, while temporary surpluses are an appropriate response to temporary favourable shocks. External balance in the face of a permanent shock, however, generally requires a rapid adjustment to current-account balance.

Similarly increases in the productivity of investment can justify a current- account deficit that is fully consistent with external balance in a long-run sense. In terms of Eq. 6, a technological innovation implying a gradual upward shift of the production function $x[k(t), t]$ generates higher levels of consumption and investment, and thus an initial current-account deficit. The ability to borrow abroad prevents the sharp rise in the interest rate that would occur initially in a closed economy; a higher investment level than under intertemporal autarky is supported by the foreign capital inflow. As productivity growth returns to normal, investment falls and current-account balance is restored with consumption and output at permanently higher levels.

These points can be made graphically in terms of a two-period Fisherian model (see Fig. 3). The axes measure amounts of the two goods available, present and future consumption, and the indifference curves show preferences over those goods. Investment opportunities are described by the production-possibilities frontier, which indicates the amount of future consumption obtained from a given input of present consumption. With the opportunity to borrow abroad at an interest rate ρ, the economy chooses to invest at point A and consume at point B, both of these points lying on the economy's budget line, which has slope $-(1 + \rho)$. Given preferences and technology, it is optimal for this economy to run a first-period current-account deficit equal to the horizontal distance between B and A; in period two, the country runs a surplus to repay its earlier borrowing. External balance thus entails an initial current-account deficit for the country shown, but surpluses for countries whose autarky interest rates are less than the equilibrium world rate ρ. The model is a parable of the development process.

When distortions in the economy cause the actual current account to diverge from its optimal level, governments may find it appropriate to adopt policies, such as taxes or subsidies on capital movement, that move the economy closer to the ideal external balance. Policies that operate directly on the distortions in question (if these can be identified) will, as usual, be best. Interesting problems arise when the countries being analysed are large enough that their governments can affect world real interest rates (and other world prices) through their actions. In this situation, the normative guidelines offered by the above approach are not directly applicable to policy analysis, and governments instead condition their actions on the conjectured responses of other governments. A Nash–Cournot equilibrium, in which each government maximizes over policy settings taking as *given* the policies of other governments, will in general be Pareto-inefficient from a global viewpoint. When governments recognize their policy interdependence, welfare in each country can be improved though policy cooperation. The practical difficulty lies in the negotiation process through which all parties agree to choose a particular point on the world contract curve.

Sovereign Borrowing and Credit Constraints

The intertemporal analysis of external balance sketched above assumes a world in which individuals, or at least governments, can borrow unlimited amounts in the world capital market, subject only to their intertemporal budget constraints.

Individual and sovereign borrowers alike, however, often appear to face binding credit constraints as a result of nonrepayment risk. After the early 1980s, the extreme difficulty for many industrializing countries of tapping world credit markets focused attention on how countries' borrowing possibilities are affected by the possibility of sovereign debt default. The problem is a central

one because most developing-country debts are either contracted directly by government agencies or are government-guaranteed.

Eaton and Gersovitz (1981) presented an early explicit analysis of the sovereign repudiation problem in an international setting. Claims on sovereign debtors are usually not legally enforceable, so the analysis of sovereign default cannot be conducted in terms of bankruptcy laws that govern cases of individual default. Eaton and Gersovitz hypothesized that a sovereign debtor defaults whenever the present discounted benefit of doing so exceeds the present discounted cost. Potential lenders, understanding the debtor's decision rule, will never lend so much that a sure incentive to default is created. Accordingly, sovereign borrowers may find themselves credit-rationed, unable to borrow as much as would normally be optimal at the interest rate quoted by lenders.

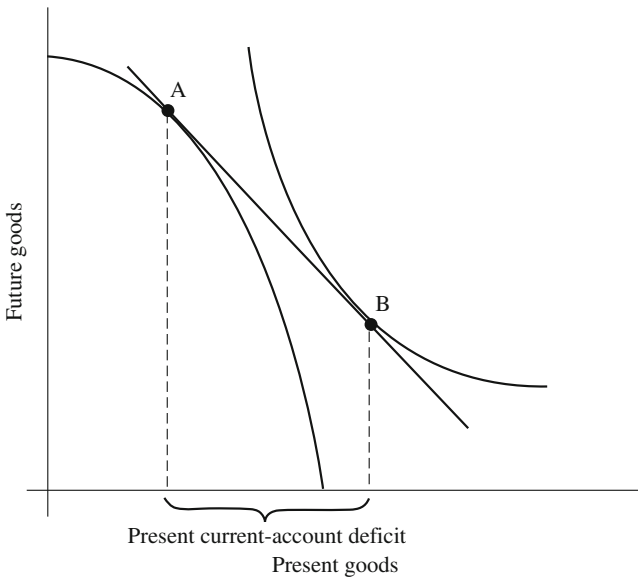
There are several potential costs of sovereign default. A defaulting country's external assets, such as foreign reserves or goods in transit, can be seized. The country could, in addition, find itself unable to borrow in the future in response to unexpected changes in its income or technology. Continued participation in the world trade and payments system might become infeasible altogether.

This 'willingness to pay' hypothesis has radical implications for the analysis of external balance. The borrowing country shown in Fig. 3, for example, would repudiate its foreign debt if that action were costless, thus avoiding the resource transfer it would otherwise have to make in the second period. As a result, period- one borrowing would take place at a country-specific interest rate reflecting the probability of default, with the extent of borrowing limited by the market's estimate of default costs. At interest rates so high that default was certain, no lending at all would occur.

The analysis of external balance becomes much more complex in such a setting.

Not only is the allowable current-account deficit more severely circumscribed; in addition, the policymaker must consider how various policy actions will affect the costs of default and hence the availability of foreign credit. Trade liberalization measures that move the economy away from an autarkic production allocation increase the cost of default by making the economy more vulnerable to disruption of its foreign trade. Such measures will therefore ease international credit constraints at the same time as they improve the static allocation of national resources. Conversely, trade restrictions aimed

International Finance, Fig. 3



at improving the current account may well reduce a country's creditworthiness.

The traditional balance-of-payments target has a rationale if the government believes that foreign credit lines may disappear unexpectedly. There is then a case for holding precautionary reserves to finance current-account deficits that may become necessary at times when credit happens to be tight or non-existent. The same purpose would be served, however, if foreign assets held by government agencies other than the central bank were run down at such times.

Internal and external balance may be irreconcilable for countries that seek to continue external debt service in the face of severe limitations on foreign borrowing. After the early 1980s, many developing countries were able to obtain private external finance only through 'forced' bank lending orchestrated by the IMF and central banks. Measures to reduce current-account deficits in line with the external funds available (and in line with IMF stabilization targets) pushed many economies into deep recession. As of this writing, it is unclear how long it will remain politically feasible for debtor governments to downplay internal-balance goals in order to continue avoiding default. There are increasingly frequent calls for some form of debt relief. Such proposals amount to the *ex post* indexation of debt contracts to adverse contingencies that were not entirely under the debtors' control.

The debt crisis of the 1980s has raised deep and consequential questions about the types of assets traded between developed and developing countries. Before the debt crisis, the typical loan contract between banks and developing-country borrowers was indexed only to the London Inter-Bank Offered Rate, and not to other factors that might alter the borrower's ability to repay. Trade between developed and developing countries in a wider spectrum of state-contingent assets would improve the international allocation of risk, and thus help to avoid future debt crises. A greater share for equity in settling current-account imbalances is one possible step in this direction. Such reforms would not eliminate the sovereign-default problem entirely, nor would they eliminate the moral-hazard problem emphasized by critics of debt-relief proposals. The possibility of a widespread and synchronized default could be sharply reduced, however, under innovative external financing arrangements.

The structure of international financial intermediation also has implications for the mutual adjustment process of industrialized countries. Current-account imbalances are only one avenue through which countries can maintain long-run consumption levels in the face of real income fluctuations or changes in investment productivity. Similar consumption-smoothing can be obtained with smaller current- account imbalances if there is a greater degree of international portfolio diversification. Lucas (1982), for example, models a world of two exchange economies with perfect international risk sharing in which consumption levels can be perfectly correlated internationally even though current-account imbalances never take place. The problem of external balance therefore never arises in Lucas's idealized setting. In reality, the extent of international portfolio diversification seem to be much smaller than plausible financial models of an integrated world capital market would predict. Why this should be so is a major empirical puzzle, and a problem for policy as well.

See Also

- ▶ Purchasing Power Parity
- ▶ Specie-Flow Mechanism

Acknowledgments *Recent general perspectives on international finance may be found in* International Monetary Fund, new open economy macroeconomics and World Bank, *as well as entries on various specific aspects of international economics*.

Bibliography

Dornbusch, R. 1973. Currency depreciation, hoarding, and relative prices. *Journal of Political Economy* 81: 893–915.

Eaton, J., and M. Gersovitz. 1981. Debt with potential repudiation: Theoretical and empirical aspects. *Review of Economic Studies* 48: 289–309.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.
Frenkel, J.A., and H.G. Johnson, eds. 1976. *The monetary approach to the balance of payments*. Toronto: University of Toronto Press.
Friedman, M. 1953. The case for flexible exchange rates. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.
Hume, D. 1752. Of the balance of trade. In *Essays moral, political and literary*, ed. D. Hume. London: Longmans Green. 1898.
Johnson, H.G. 1969. The case for flexible exchange rates, 1969. *Federal Reserve Bank of St. Louis Monthly Review* 51 (6): 12–24.
Keynes, J.M. 1936. *The general theory of employment interest and money*. London: Macmillan.
Lucas, R.E. Jr. 1976. Econometric policy evaluation: A critique. In *The Phillips curve and labor markets*, ed. K. Brunner and A.H. Meltzer. Amsterdam: North- Holland.
Lucas, R.E. Jr. 1982. Interest rates and currency prices in a two-country world. *Journal of Monetary Economics* 10: 335–359.
Meade, J.E. 1951. *The balance of payments*. London: Oxford University Press.
Metzler, L.A. 1948. The theory of international trade. In *A survey of contemporary economics*, ed. H.S. Ellis. Philadelphia: Blakiston.
Metzler, L.A. 1960. The process of international adjustment under conditions of full employment: A Keynesian view. In *Readings in international economics*, ed. R.E. Caves and H.G. Johnson. Homewood: Irwin. 1968.
Mundell, R.A. 1961. The international disequilibrium system. In *International economics*, ed. R.A. Mundell. New York: Macmillan. 1968.
Mundell, R.A. 1968. The nature of policy choices. In *International economics*, ed. R.A. Mundell. New York: Macmillan.
Nurkse, R. 1944. *International currency experience: Lessons of the inter-war period*. Geneva: League of Nations.
Obstfeld, M. 1985. Floating exchange rates: Experience and prospects. *Brookings Papers on Economic Activity* 1985 (2): 369–464.
Obstfeld, M., and A.C. Stockman. 1985. Exchange-rate dynamics. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen. Amsterdam: North- Holland.
Ricardo, D. 1817. *The principles of political economy and taxation*. London: J.M. Dent & Sons. 1911.
Samuelson, P.A. 1971. On the trail of conventional beliefs about the transfer problem. In *Trade, balance of payments, and growth*, ed. J.N. Bhagwati et al. Amsterdam: North-Holland.
Triffin, R. 1960. *Gold and the dollar crisis*. New Haven: Yale University Press.
Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper & Brothers.

International Financial Institutions (IFIs)

John Toye

Abstract

The International Monetary Fund was established to manage international payments post-Second World War. The gold exchange standard re-established current account convertibility in the industrialized nations and oversaw rapid growth of international trade. After that standard collapsed in 1971 the IMF ran stabilization programmes for developing countries, with mixed success. The World Bank was set up to provide medium-term loans at concessional interest rates for (post-war) reconstruction and to develop capital-poor areas. In 1979 it initiated programme lending with conditions to promote economic adjustment. Conditionality has been under-enforced but increasingly loans go to countries that show commitment to liberal economic reforms.

Keywords

Baker Plan; Bretton Woods agreement; Capital account controls; Conditionality; Convertibility; Fixed exchange rates; Floating exchange rates; Gold exchange standard; International capital flows; International Development Agency (IDA); International financial institutions; International Monetary Fund; Keynes, J.M.; Mexican debt crisis 1982; Poverty Reduction Strategy Papers; Special Drawing Rights (SDRs); World Bank

JEL Classifications

O16

The two major international financial institutions, the International Monetary Fund (IMF) and the World Bank, were designed on an Anglo-American plan, negotiated by John Maynard Keynes and Harry Dexter White. After gaining wider approval

at the Bretton Woods conference in July 1944, they were set up in order to introduce new elements of multilateral regulation into the working of the international economy (Skidelsky, 2000)

The International Monetary Fund

The IMF, established in 1947 to manage international payments in the economic chaos that followed the Second World War, had four Charter objectives: to restore a system of multilateral payments for current transactions between its members; to minimize disequilbrium in the members' international balances of payments; to promote exchange stability; and 'to facilitate the expansion and balanced growth of international trade, and to contribute thereby to the promotion and maintenance of high levels of employment and real income'. In promoting all of these objectives, the Fund originally acted as the umpire of a set of rules of international monetary behaviour.

Originally, the IMF managed a system of fixed, but adjustable, exchange rates against the US dollar, which itself was pegged to gold. In order to keep exchange rate fluctuations within set limits, each member country – and the membership then was much smaller than it was in 2005 (over 180) – paid into the Fund a capital sum, determined according to its importance in world trade, and was given a borrowing 'quota' related to its capital. Voting power in the organization is related to the size of this capital. In balance of payments difficulties, members were permitted to borrow from the Fund and repay over the following two or three years. Thus the Fund acted as a bank, but the scale of the 'banking' operation was initially small. Between 1947 and 1955, 14 out of 59 members made drawings, at an annual rate of $46 m. This equalled 0.06 per cent of world imports. In 1990–8, when 78 out of 182 members made drawings, the rate was $13.4bn, or 0.29 per cent of world imports.

The gold exchange standard succeeded in re-establishing current account convertibility in the industrialized nations, while permitting countries to maintain capital account controls. The IMF had less success in shortening and reducing the severity of balance of payments disequilibria (Killick, 1985). Nevertheless, under this system, international trade did grow rapidly, and employment and real income also grew faster than subsequently. Fear of liquidity shortage led the IMF in 1967 to create Special Drawing Rights (SDRs) – the First Amendment of the Fund's Articles of Agreement – but they came too late to save the anchor of the system, the $35 an ounce fixed parity of the official price of gold. The ratio of US gold reserves to its liquid liabilities had fallen from 2.73 in 1950 to 0.41 by 1968. Once the private market gold price rose above the official price, dollar-gold convertibility was suspended de facto, and officially abandoned in 1971. The collapse of the gold exchange system was thus due to an inherent design flaw, and not to any particular failures of the IMF.

The Fund soon ceased to be a banker to OECD countries, and began to cast around for a new role in the developing countries. However, this changed the Fund from an institution of collective action for industrial countries into their instrument for disciplining others. The Second Amendment to the IMF Articles in 1978 allowed all forms of national exchange-rate mechanism, except pegging to gold. Many larger economies chose to float their currency. Many smaller economies chose to peg their exchange rate to other currencies or baskets of currencies. Systemic international economic coordination was replaced by G7 meetings that tried to 'talk down' or 'talk up' particular key currencies.

Under the gold exchange standard, developing countries had been of little interest to the IMF. Many had never been properly integrated into the system, although in Peru and Paraguay the Fund did pioneer policy-conditioned lending. From the early 1960s, under UN pressure, the Fund developed additional 'banking' facilities relevant to the needs of developing countries – for example, the Compensatory Financing Facility in 1963 and the Extended Fund Facility (EFF) in 1974, which provided medium-term finance, beyond the limits of normal lending, to support agreed stabilization programmes requiring structural adjustment.

The Mexican debt crisis of 1982 was a turning point in the history of the Fund. Following the

Baker Plan of 1985, the US Administration recruited the Fund, along with the World Bank, to be its managers at one remove of the prolonged debt crisis that for some years threatened the survival of major Western banks. The capital available to both institutions was increased. Building on the EFF, new longer-term lending facilities were created to channel credit to indebted developing countries. In 1986 the Structural Adjustment Facility was set up, and in 1987 the Extended Structural Adjustment Facility (ESAF), to provide loans to low-income countries suffering protracted balance of payments problems at 0.5 per cent interest over five and a half to ten years. Policy conditionality is strong under ESAF loans, and is specified in the Poverty Reduction Strategy Papers of the borrowing country.

However, IMF stabilization programmes frequently broke down before completion. Between 1979 and 1993, 53 per cent of 305 Fund programmes were uncompleted, often because of inadequate financing (Killick, 1995, pp. 58–65). If sustained, they improved the current account and the overall balance of payments, and slowed inflation, but at the cost of a short-term reduction in growth. The Fund came under particularly fierce criticism for its handling of the Asian financial crisis of 1997–8 (Stiglitz, 2002). It has since introduced reforms to improve its own transparency and member countries' data reporting standards.

The World Bank Group

The International Bank for Reconstruction and Development (IBRD) was established in 1946 in order to provide medium-term loans at less than commercial interest rates to governments for (post-war) reconstruction and for the development of capital-poor areas. Since then other parts of what is now called the World Bank Group have been added – the International Finance Corporation (IFC), set up in 1956 for lending to the private sector, the International Centre for the Settlement of Investment Disputes (1966) and the Multilateral Investment Guarantee Agency (MIGA) (1988). However, the most significant addition was the International Development Agency (IDA) in 1960, to provide long-term, highly concessional loans to the poorest countries.

Having largely missed out on post-war reconstruction lending, the Bank focused on project lending for economic development. Its procedure was to borrow on the developed country capital markets and re-lend (plus a small margin) for specific investment projects in developing countries. In the early years, this was a slow process, originally concerned with large physical infrastructure schemes, such as dams and electricity generation. After IDA gave the Group a development agency function, the composition of Bank investments began to change, gradually including agricultural and urban redevelopment projects. The criterion of project success was the *ex post* rate of return on each project. In 1973, a semi-independent Operations Evaluation Department was established to calculate this. The Bank's participation almost certainly produced a better quality of project than would have occurred in its absence. However, if fungibility exists, the economic effect of the investment cannot be measured by its *ex post* rate of return. Although fungibility need not concern a development bank, whose chief aim is to recover its loans, it should worry a multilateral aid agency funded by public capital, whose main objective is to promote the sound development of the borrower's economy. To ensure that, projects need to be appraised as part of a comprehensive development plan.

In the 1970s, when the economies of developing countries were disturbed by substantial economic shocks, the World Bank decided that the success of their individual loan projects, as measured by their *ex post* rates of return, was being affected negatively by their broader economic environment (rising oil price, high inflation, fixed nominal exchange rates, import restrictions, and so on). In 1979, the Bank initiated programme lending, previously regarded as an unsound banking practice. The new types of loans, structural (SAL) and sectoral (SECAL) adjustment lending, provided rapidly disbursing foreign exchange on condition that the borrowing government undertook economic policy changes, either economy-wide or sectorally.

Programme lending with policy conditions attached provided the instrument that the Bank could bring to the task of co-managing the 1980s debt crisis with the Fund. A Fund–Bank 'concordat' in 1989 established effective (though not formal) cross-conditionality of Fund and Bank loans. Bank adjustment lending became conditional on a pre-existing Fund programme, and a statement of economic policy for the borrowing country had to be agreed by both institutions – entitled the Poverty Reduction Strategy Paper.

The evaluation of the effects of programme lending is more difficult and controversial. Governments of developing countries have been reluctant to comply with some of the conditions for policy change laid down in the loan agreements. This is often described as a result of their 'lack of ownership' of the economic reform process. The Bank itself faces incentives that make it unlikely that it will react to non-compliance consistently with a discontinuation of funding (Mosley, Harrigan and Toye, 1995). Thus the evidence suggests that the Bank's loan conditionality is a weak instrument for inducing policy change (see Ferreira and Keely, 2000). At the start of the 21st century, the Bank was moving towards a lending strategy of selectivity, in which future loans are directed increasingly to countries that have already demonstrated their zeal for neoliberal economic reform.

The Bank has been criticized on the grounds that private flows to developing countries can do the job instead (Krueger, 1998). In 1970, IBRD net lending was about ten per cent of net private flows. In 1996, this share had fallen to 0.7 per cent. In 25 years, private flows had increased 40-fold, while IBRD flows had increased threefold in nominal terms. The original justification of IBRD loans in terms of imperfect private capital markets seems weak in the light of these figures, although private finance is very concentrated geographically and the Asian crisis showed how short term and volatile private money can be.

Apart from lending, the Bank undertakes many other activities. It conducts what is probably the largest single publication programme on development issues in the world. This includes its own research across the field of development problems, published in two house journals, flagship reports like the annual *World Development Report*, a host of monographs and a multitude of Working Papers. The Bank has also become a major provider of statistical data, including regular published series and data from household and firm surveys. It regards itself as a 'knowledge agency'.

See Also

- ▶ Foreign aid
- ▶ Keynes, John Maynard (1883–1946)
- ▶ Third world debt

Bibliography

Akyuz, Y. 2002. *Reforming the global financial architecture: Issues and proposals*. Geneva: UNCTAD.

Boughton, J. 2001. *Silent revolution: The international monetary fund, 1979–89*. Washington, DC: IMF.

Buira, A., ed. 2005. *The IMF and the World Bank at Sixty*. London: Anthem Press.

Ferreira, F., and L. Keely. 2000. The World Bank and structural adjustment: lessons from the 1980s. In *The World Bank: Structure and policies*, ed. C. Gilbert and D. Vines. Cambridge: Cambridge University Press.

Kapur, D., J. Lewis, and R. Webb. 1997. *The World Bank: Its first half century. volume 1: History*. Washington, DC: Brookings Institution.

Killick, T. 1985. *The quest for economic stabilization: The IMF and the Third World*. Aldershot: Gower.

Killick, T. 1995. *IMF programmes in developing countries: Design and impact*. London: Routledge.

Krueger, A. 1998. Whither the World Bank and the IMF? *Journal of Economic Literature* 36: 1983–2020.

Mallaby, S. 2005. Saving the World Bank. *Foreign Affairs* 84(3): 75–85.

Mosley, P., J. Harrigan, and J. Toye. 1995. *Aid and power: The World Bank and policy-based lending*. Vol. 1. 2nd ed. London: Routledge.

Ritzen, J. 2005. *A chance for the World Bank*. London: Anthem Press.

Skidelsky, R. 2000. *John Maynard Keynes: Fighting for Britain 1937–1946*. London: Macmillan.

Stiglitz, J. 2002. *Globalization and its discontents*. London: Allen Lane.

Toye, J., and R. Toye. 2004. *The UN and global political economy: Trade, finance and development*. Bloomington: Indiana University Press.

Woods, N. 2001. Making the IMF and the World Bank more accountable. *International Affairs* 77: 83–100.

International Income Comparisons

Irving B. Kravis

Despite the accumulating evidence of systematic error, the most common means of making international comparisons still remains the conversion of incomes expressed in own-currencies to a numeraire currency, most often the US dollar, via the exchange rate. The procedure has long been held suspect by travellers who could observe that some countries were dear and others inexpensive – i.e. that exchange rates did not reflect the purchasing power of currencies. As will be shown, the purchasing power of the currencies of poor countries tends to be understated by exchange rates. Exchange-rate conversions thus tend to exaggerate the dispersion of the real per capita incomes of the different nations.

Hence the unique character of making valid *international* income comparisons arises from the existence of different currency units and the need to compare their purchasing powers. For the most part, the other conceptual and empirical problems of international income comparisons are similar to those of within-nation comparisons across time or between persons at a given period.

Ignoring index number problems, the basic approach in recent international comparisons may be simply described as involving the derivation of a quantity (real income) comparison by dividing a price ratio into an expenditure ratio:

$$\frac{Q_j}{Q_b} = \frac{E_j}{E_b} \div \frac{P_j}{P_b}$$

where j and b are countries; Q's are physical quantities, E's are expenditures (for GDP or its components), and the P's are prices, the E's and P's being in own currencies. (P_j/P_b is the purchasing power parity.) Why not, it may be asked, make direct quantity comparisons of the commodities and services that make up the real incomes being compared? The answer is that it is more difficult for most kinds of goods to get a representative sample of country-to-country quantity ratios for the same or equivalent goods than it is to get a representative sample of price ratios. Also, the quantity ratios are more likely to be subject to greater sampling variations than price ratios.

The Evolving System of Comparisons

The history of international income comparisons includes many ad hoc efforts based on exchange rate conversions and only a few careful attempts to compare the purchasing power of currencies (Kravis 1984). The system of international income comparisons that is emerging towards the end of the 20th century has its origins in the study of the Organization for European Economic Cooperation (OEEC) by Gilbert and Kravis (1954). The OEEC study laid down a pattern that has been followed in the UN International Comparison Project (ICP) and related studies carried out by regional groups. In particular, the use of the price-times-quantity-equals-expenditure relationship and the breakdown of GDP in terms of its final product components rather than in terms of producing industries were carried over into the ICP work.

As of early 1986 the system included official benchmark studies for 60 countries with a 1980 reference date, and for diminishing numbers, for earlier reference dates as well, as far back as 1967 (UN and EC 1986; Ward 1985; Kravis, Heston, and Summers [hereafter KHS] 1982). Unofficial estimates based on extrapolations, to be described presently, were available for most other countries (Beckerman 1966; KHS 1978b; Summers and Heston 1984).

The evolution of the system has been greatly influenced by two major developments that had pervasive effects on the statistical description of the world economy. One was the emergence of national income accounting beginning in the years preceding World War II. Under the aegis of the United Nations a standardized system of national accounts was developed that was adopted by most nations of the world. The standardized system provided a common statistical framework for

international comparisons with respect to such important matters as the definition of income (the gross domestic product concept) and the prices at which goods are to be valued (producers' market prices). Earlier international income comparisons tended to focus on the incomes of selected groups of wage earners or employees and to exclude incomes other than those arising from employment. The recent ones are therefore more comprehensive in scope, both with respect to the types of income and population coverage than the past wage-earner-oriented studies. The fact that the ICP studies offer separate estimates for personal consumption may offset the disadvantage that they are based on the inclusion of gross capital formation rather than on some net concept.

The other major development, the advent of the computer, has greatly expanded the availability of methods that meet the needs of comparisons involving many countries at once. In most studies preceding those of the ICP, small numbers of countries were involved and the method turned on a series of binary comparisons – i.e. comparisons between pairs of countries – sometimes, as in the OEEC studies, with all the other countries compared with a country selected as the centre country (star system). This approach has the advantages of simplicity and ease of understanding. An important disadvantage is that the quantitative relationships among the other countries, derived from the relation of each to the centre country, will vary with the choice of a country for the central role. Without the computer, it would have been infeasible to find and apply methods that were invariant to the selection of the base-country, and which at the same time had other desired properties. An important additional property for the real income comparisons that was sought and attained is matrix consistency. This property is akin to that afforded by a time-to-time national accounts table showing the income originating in different economic sectors in constant prices. That is, the figures in any one column (pertaining to a given country) may be added to yield aggregate GDP and subaggregates (e.g. consumption, food, etc.), and the figures on any row (each pertaining to a final expenditure category) show the correct quantity relationships between the different countries. (See, for example, KHS 1982, p. 19.)

The Actual Work of Preparing the Comparisons

The actual work of the international income comparisons consists mainly of making price comparisons. The tasks involved are:

(1) Subdividing GDP into categories for which expenditure data and price comparisons can be obtained.
(2) Selecting and pricing a sample of specifications for each expenditure category.
(3) Aggregating the price relatives at the category level.
(4) Aggregation of the categories to form price and quantity indexes for GDP and its subaggregates.

In the literature on international income comparisons, the most intellectual effort has gone into devising index number formulas for the aggregation of the categories, a problem that has long fascinated economic statisticians. The formula chosen in the ICP is one which in principle values the quantities of goods in each country's GDP at world average prices. These values when summed yield the desired real income comparisons or comparisons for components of real income (e.g. consumption, food, beef). The formula, which was suggested by Robert Geary and amplified by S. Khamis, involves deriving the price comparisons and the average world prices simultaneously in two subsets of equations. Some statistical experiments have suggested that radically different results would not be produced by alternative formulas which have equally plausible claims to consideration (KHS 1982, pp. 95ff).

It has been objected that the per capita quantity indexes for poor countries are inflated because the weights are dominated by the expenditure of the rich countries. However, some experimental work indicates that the results are not greatly changed when the weights of low income countries are

greatly increased in calculating the world average prices (Kravis 1984).

The quality of the income comparisons is, as a practical matter, more vulnerable to the care with which the price comparisons are carried out than to any other phase of the work. Not only must the sample of specifications of each detailed category be representative of price formation influences in each country but the items actually priced in the different countries must be equivalent in quality. Among the means used in the ICP to ensure such equivalents were international exchanges of samples, visits by price experts from one country to partner countries to consult with their counterparts and to examine goods in shops, and resort to informal and formal advice of merchants, manufacturers and engineers. Once specifications were identified, it was necessary to ensure that the price provided was the national average price; this was the responsibility of the country's statistical authorities to supply, sometimes from prices collected for other purposes and in other cases from special price surveys.

Certain services for which outputs are difficult to measure, including education, medical care and government, cannot always be treated in this standard specification approach. In some cases, as services of physicians, quality-adjusted inputs were used as proxies for measures for the international comparison of outputs. Some have claimed that the treatment of these services led to an overstatement of the real per capita GDP of low income countries in Phase III (Maddison 1983) but sensitivity analysis indicates the possible impacts on real GDP per capitas of different treatments are small (Kravis 1986).

Substantive Findings

Comparisons based both on PPP and exchange rate conversions are shown in Table 1 for a selected set of countries. The countries are arranged by region and within region by ascending order of real GDP per capita; the exchange rate deviation index, the ratio of the PPP to the exchange rate conversion, is shown in column 3. It can be seen from column 3 that the estimates based on the PPP conversions tend to be higher for lower income countries. That is, poor countries tend to have lower price levels; their exchange rates understate the purchasing power of their currency. A consequence is that the spreads between the average incomes of the countries greatly diminishes when PPP conversions are used. For example, the ratio of the highest to the lowest per capita on the exchange rate basis is nearly 100 to 1 (Germany to Ethiopia) whereas on a PPP basis it is a little over 40 to 1 (Canada to Ethiopia).

Two lines of explanation have been offered for the tendency towards low prices in poor countries. In the productivity differential model, the productivity of poor countries is held to be lower in both traded and nontraded goods but by smaller differentials in nontraded goods (e.g. teaching). Prices of

International Income Comparisons, Table 1 Real per capital GDP, selected countries, 1980

GDP per capita	Converted by PPP (1)	Exchange rate (2)	Exchange rate – deviation index (3) = (1) ÷ (2)
Africa			
Ethiopia	2.5	1.2	2.1
Kenya	5.6	3.7	1.5
Ivory Coast	12.0	11.2	1.1
Asia			
India	5.0	2.1	2.4
Korea	22.6	13.3	1.7
Japan	73.5	77.8	0.9
Europe			
Portugal	33.5	21.2	1.6
Spain	55.5	49.2	1.1
Italy	68.0	60.3	1.1
UK	72.1	81.6	0.9
France	85.4	106.0	0.8
Germany	89.1	116.2	0.8
Latin America			
Brazil	29.3	18.0	1.6
Argentia	33.6	47.4	0.7
North America			
Canada	101.5	94.5	1.1
US	100.0	100.0	1.0

Sources: UN and Commission of the European Communities (1986)

I

traded goods tend to be drawn to international levels; low productivity in a poor country thus means low wages. However, the same wage level will also prevail in the poor country's nontradables sector, but with productivity somewhat better, prices will be lower. An alternative explanation turns on the labour-abundant factor endowments of poor countries and assumes that nontraded goods (especially services) are labour-intensive and therefore cheap in poor countries. The average price level is pulled down not only by low prices for nontradables but because tradables too, are cheaper since they almost always are sold with nontradable components (e.g. distribution costs).

Aside from the light shed on real per capita income, the ICP studies illuminate two other important aspects of the world economy. One is that it provides a comparison of the general (GDP) price level of the different countries. The price level is the ratio of the PPP to the exchange rate; it is the reciprocal of the exchange rate deviation index. Wide variations in price levels can be seen to exist even between different members of the European Common Market. These measures add to the existing information on *relative* movements of price levels, a measure of the *absolute* gap. They indicate for example that the German price level was 130 per cent of that of the US in 1980 and only 76 per cent of the US level in 1984 (Ward 1985).

The other broad set of insights into the structure of the world economy arises from the price and quantity comparisons that are available for components of GDP. A host of questions, many arising in connection with basic economic analyses such as cross-country demand studies can be answered by these comparative price and quantity data. How do food prices differ in low and high income countries? The quantity of medical care? The extent of R&D in real terms? The amount of government services? and so on.

The Future of International Income Comparisons

Since it is clear that benchmark estimates will not soon be available for many more than 60 to 70 countries, some less costly even if more approximate method of estimating real per capita GDP will have to be employed for the remaining 50 or so non-benchmark countries. Several approaches have been tried (Kravis 1984, p. 18). One approach is based on an estimating equation that embodies the relationship between real GDP per capita of benchmark countries and certain physical indicators such as milk or steel consumption (Beckerman 1966). Another uses certain widely available macroeconomic variables for the extrapolating equation (KHS 1978a; Summers and Heston 1984). (For example, real GDP per capita is taken as a function of exchange-rate-converted GDP per capita and the propensity to trade as measured by the ratio of exports plus imports to GDP.) The statistical margins of error surrounding these 'shortcut' estimates for non-benchmark countries make explicit the degree of uncertainty, in contrast to the seemingly unambiguous estimates produced by the exchange rate conversions. However, the exchange rate conversions are not error free; they are known to be biased. In fact, the shortcut estimates come closer than the exchange rage conversions to what full benchmark studies would yield, much closer for low and middle income countries (Kravis 1986). Benchmark studies would be best, but given that they will not be available for all countries in the near future, a mixed set of benchmark and short cut PPP estimates should be used.

See Also

- National income
- Purchasing power parity
- Real income

Bibliography

Beckerman, W. 1966. *International comparisons of real income*. Paris: OECD Development Centre.

Gilbert, M., and I.B. Kravis. 1954. *An international comparison of national products and the purchasing power of currencies: A study of the United States, the United Kingdom, France, Germany, and Italy*. Paris: OEEC.

Kravis, I.B. 1984. Comparative studies of national incomes and prices. *Journal of Economic Literature* 22(March): 1–39.

Kravis, I.B. 1986. The three faces of the international comparison project. *World Bank Research Observer* 1(January): 3–26.
Kravis, I.B., A.W. Heston, and R. Summers. 1978a. *International comparisons of real product and purchasing power*. Baltimore: Johns Hopkins University Press.
Kravis, I.B., A.W. Heston, and R. Summers. 1978b. Real per capita for more than one hundred countries. *Economic Journal* 88(June): 215–242.
Kravis, I.B., A.W. Heston, and R. Summers. 1982. *World product and income: International comparisons of real gross product*. Baltimore: Johns Hopkins University Press.
Maddison, A. 1970. *Economic progress and policy in developing countries*. New York: Norton.
Maddison, A. 1983. A comparison of the levels of GDP per capita in developed and developing countries, 1790–1980. *Journal of Economic History* 43(March): 27–41.
SOEC (Statistical Office of the European Community). 1977. *Comparisons in real values of the aggregates of ESA, 1975*. Luxembourg: EEC.
SOEC (Statistical Office of the European Community). 1983. *Comparison in real values of the aggregates of ESA, 1980*. Luxembourg: EEC.
SOEC (Statistical Office of the European Community). 1985. *Comparison of price levels and economic aggregates: The results for African countries*. Luxembourg: EEC.
Summers, R., and A.W. Heston. 1984. Improved international comparisons of real product and its composition: 1950–80. *Review of Income and Wealth* 30(2): 207–262.
United Nations. 1985. *National accounts statistics, main aggregates and detailed tables, 1982*. New York: United Nations.
United Nations and Commission of the European Communities. 1986. *World comparisons of purchasing powers and real product for 1980: Phase IV of the international comparison project: Part I: Summary results for 60 countries*. New York: United Nations.
Ward, M. 1985. *Purchasing power parities and real expenditures in the OECD*. Paris: OECD.

International Indebtedness

Vladimir Brailovsky

After World War II many less-developed economies started industrialization programmes which contributed to the achievement of rapid and sustained growth of income. Although industrialization meant a continuous reduction of the import propensity of their economies, growth also implied that the level of imports tended to exceed that of exports. This gap was covered by external indebtedness and, to a lesser extent, by direct foreign investment. During the 1950s and the early 1960s, credit was granted mainly by the governments of advanced economies and multilateral financial organizations. Starting from the mid-1960s, however, international indebtedness was increasingly dominated by private banking, reducing the element of aid implicit in previous arrangements. More importantly, this shift in the nature of credit flows was less conducive to the coordination of policies between industrial and developing countries, the consequences of which became apparent later on.

The remarkable expansion of the world economy achieved during the 1960s was suddenly interrupted at the beginning of the 1970s, due mainly to the policy responses of advanced Western economies to the increase in commodity prices, especially oil. The dramatic increase in oil prices in 1973–4 and again in 1979–80 would have implied, *ceteris paribus,* a situation in which -current-account surpluses of OPEC countries and other producers would have been mainly reflected in compensating deficits of the OECD economies – the major importers of oil – and to a lesser extent in deficits of non-oil producers in the Third World. In fact, however, the deficits of the latter swelled much more than what would have been expected on the basis of this assumption. The reason being that, in order to eliminate their own deficits, the advanced economies applied restrictive fiscal and monetary policies which reduced their growth rates (Llewellyn et al. 1985). This affected disproportionately the exports from developing economies, thereby increasing their need for foreign lending. Private banking was instrumental in 'recycling' large amounts of petro-dollars – indirectly, through financial intermediation – into the debit side of the balance-sheets of these economies.

Thus, between 1973 and 1980, the OECD economy was only required to shift real resources to oil producers in the form of extra exports – the only inevitable cost arising from oil price

increases – of the order of half a percentage point of GDP. The rest of the oil earnings was used to accumulate financial assests. However, in spite of the small magnitude of this real transfer, the annual growth rate of income in these economies dropped two percentage points below its long-term trend. By 1980 the loss accumulated to 15 per cent of GDP (CEPG 1980).

The maintenance of growth in many underdeveloped economies during the 1970s, plus the availability of international finance in great amounts created by the oil surpluses, increased enormously the level of their foreign debt. The dramatic rise in interest rates at the end of this period, which took place in creditor countries as a consequence of the application of monetarist policies, compounded the problems of debt servicing. Uncoordinated policies of the international banking community made this increase in debt possible, in spite of the heavy exposure of private institutions to sovereign borrowers in the Third World. Equally irrational was their reaction at the beginning of the 1980s, when general economic conditions deteriorated and a sudden awareness of the risks involved was regained. From lending in almost limitless quantities, abruptly the banks decided not to lend at all.

It is of some interest to analyse in more detail the costs of the adjustment process that this sudden change in the availability of finance implied for developing economies. Take the national accounts identity

$$Q \equiv D + X - M \tag{1}$$

where Q is gross domestic product, D is domestic demand, and X and M are respectively exports and imports of goods and non-factorial services. The balance-of-payments identity can be represented as

$$B \equiv M - X + N \tag{2}$$

where B is net foreign borrowing less capital movements abroad and N net interest payments. If m is the import propensity, then for simplicity one can assume that

$$M = m \cdot Q \tag{3}$$

Taken together, these formulations imply that if B, X and N are given either by external circumstances or by history, then necessarily

$$D = [(B - N)(1 + m) + X]/m \tag{4}$$

$$Q = (B - N + X)/m \tag{5}$$

Under these assumptions, different mechanisms will be in operation in order to ensure that domestic demand and output attain the above values. Finally, transfers abroad of real resources are

$$T \equiv Q - D \equiv X - M \equiv -(B - N) \tag{6}$$

Take now a typical situation before the debt crisis in which $B > N$; that is, in which transfers of resources were obtained by developing countries, say by $A = B - N > 0$. Then compare it with one where $B = 0$, prevailing after the crisis. This quantity can even be negative if capital flights increase, as is usually the case when doubts are cast about the creditworthiness of a country. Whereas before the economy was receiving A, now it is transferring resources abroad equivalent to the amount of N. The difference is $(A + N)$. More importantly, the difference in domestic demand is $(A + N)(1 + m)/m$ and $(A + N)/m$ for output. Since m is normally between 0.1 and 0.2, even if it is considerably reduced in the process of adjustment (through devaluation and other policies), this implies that the economy is not only transferring $(A + N)$, but that in order to be able to pay, domestic output and demand must drop by a large multiple of $(A + N)$. Whereas the former are resources obtained by creditor nations, the latter are simply wasted for the world economy as a whole. If this adjustment process takes place simultaneously in several economies, the wastage is even greater since X will also tend to fall.

Once this basic relationship between transfers abroad and the domestic levels of demand and output is grasped, it is straightforward to understand the nature of the policy recipes of international organizations such as the IMF. They are directed mainly towards the lowering of domestic

demand through cuts in both the public sector deficit and consumption, the latter via a reduction of the real value of wages in terms of the exchange rate (i.e. devaluations). It is therefore not surprising that this adjustment process is normally accompanied by accelerating inflation, as a reflection of competing demands for shares in income, the level of which is drastically reduced. The external vulnerability of these economies is further enhanced in the long-rung, and capacity to pay curtailed, since prolonged situations of depression are not conducive to capital formation and productivity growth. This makes it more difficult to reduce the import propensity and to stimulate exports. The import propensity is also likely to increase if the economy follows a trade liberalization policy, another element in the book of orthodox recipes.

The real resources which less-developed economies have been transferring abroad since 1982 is considerably greater in relative terms than the amounts involved for advanced countries following the oil crisis. For example, Argentina, Brazil and Mexico have had to expand in recent years their trade surpluses to around 6 per cent of GDP (ECLA 1985). As a point of reference, Japan's trade surplus was not, at the time, greater than 2.5 per cent of output. Under these circumstances it is worth asking whether a policy, such as that proposed by the IMF and the international banks – which requires relatively poor countries to transfer abroad, year after year, a large proportion of resources, with a widening gap between potential and actual output – can be described as a permanent solution to the debt problem.

More likely than not, this situation will at some stage lead to an outright default by major debtors unless there is a change in the rules of the game. Defaults such as these have occurred in the past, and with little consequences to the debtors (Winkler 1933; Wynne 1951; Wood 1980). They may become an attractive option given the magnitude of the resources which can be reclaimed for domestic uses. Due to their heavy exposure to sovereign nations, banks have a lot to lose in this event, and there is little they can do in terms of legal and other sanctions (Kaletsky 1985). The costs of default can be minimized only to the extent that governments in advanced countries – once again, as during the 1950s and 1960s – play an active role in this field and the debt burden is shared equitably among the different parties involved. The incentive to do this lies in the fact that, given the private origin of international indebtedness, their own financial stability may be jeopardized.

See Also

- ▶ External Debt
- ▶ Fiscal and Monetary Policies in Developing Countries
- ▶ International Liquidity

Bibliography

CEPG (Cambridge Economic Policy Group). 1980. World trade and finance: prospects for the 1980s. *Cambridge Economic Policy Review* 63, Department of Applied Economics, University of Cambridge and Gower Publishing Company.

ECLA (Economic Commission for Latin America). 1985. *Estudio Económico de América Latina y el Caribe, 1984.* United Nations, Economic and Social Council, LC/L.330.

Kaletsky, A. 1985. *The costs of default*. New York: Priority Press.

Llewellyn, J., S. Potter, and L. Samuelson. 1985. *Economic forecasting and policy: The international dimension*. London: Routledge & Kegan Paul.

Winkler, M. 1933. *Foreign bonds: An autopsy.* Philadelphia: R. Swain.

Wood, P. 1980. *The law and practice of international finance*. London: Sweet and Maxwell.

Wynne, W. 1951. *State insolvency and foreign bondholders: Case histories*, vol. 2. New Haven: Yale University Press.

International Liquidity

A. D. Crockett

International liquidity may be defined as that stock of assets which is available to a country's monetary authorities to cover payments imbalances (when the exchange rate is fixed) or to

influence the exchange value of the currency (when the exchange rate is flexible). A distinction may be drawn between unconditional liquidity, which is generally owned by the country concerned and may be used at its sole discretion, and conditional liquidity, which comprises access to borrowing facilities and is generally available only on conditions set by the lenders. Because of the obvious practical difficulties in measuring conditional liquidity, the operational measure of international liquidity that is generally used in discussion of the subject is that of gross international reserves.

The definition of international reserves used by the International Monetary Fund in compiling *International Financial Statistics* includes: gold; short-term foreign exchange holdings in convertible currencies; special drawing rights (SDRs); and reserve positions in the International Monetary Fund; As of December 1984, total holdings of reserves reported by member countries of the IMF amounted to SDR 438 billion (with gold valued at SDR 35 per ounce). Of this total, 8 per cent was represented by gold holdings (at SDR 35 per ounce); 10 per cent was reserve positions in the Fund; 4 per cent was SDRs; and the remainder was in holdings of foreign currencies (about 70 per cent of which was US dollars).

From an economic point of view, the most significant aspect of the subject of international liquidity is the relationship between the stock of reserves (whether for a country or the world as a whole) and other economic variables, such as the level of real output, the price level, and the pattern of balance of payments positions. It has been recognized that this relationship depends on the nature of the demand function for reserves, and of the arrangements governing reserve supply. Much of the literature on international liquidity has thus focused on these two aspects.

Concerning demand, the demand for reserves by countries, like that for national money by individual economic agents, rests on a desire to enhance welfare by cushioning fluctuations in absorption that might otherwise be made necessary by the non-synchronous nature of payments and receipts. A stock of reserves represents purchasing power that can be used to moderate the domestic economic impact of declines in foreign exchange receipts. Even where official reserve holdings are not in fact used for this purpose, their existence may facilitate the activation of international credits that serve the same purpose. Standard utility theory teaches that the optimum stock of reserves for a country will be that quantity at which the benefits of an additional unit of reserves (in terms of the flexibility it affords the monetary authorities) just balances the cost of acquiring and holding it. Key factors determining the demand for reserves are the amplitude of fluctuations to which an economy is subject in its external position; and the availability of alternative means of financing, or adjusting to, these payments disturbances. For an individual economy, therefore, reserve demand will depend on the structure of its balance of payments. Heavy dependence on exports of primary products subject to volatile supply and demand conditions will tend, *ceteris paribus*, to lead to a greater need for reserves. On the other hand, access to borrowing facilities to cover payments imbalances, or a willingness to allow the exchange rate or domestic policies to adapt so as to encourage accommodating capital flows, will reduce the demand for owned reserves.

For the world as a whole, it has generally been thought that payments disturbances would tend to grow with the underlying volume of world trade; however, there is less agreement about whether the relevant elasticity should, for practical purposes, be regarded as unity. (This debate parallels that on the income elasticity of the transactions demand for cash within a national economy). A further important issue is the extent to which exchange rate flexibility alters (presumably reduces) the demand for reserves. While governments have always had alternatives to reserve use for financing payments disequilibria (e.g. official borrowing, or the manipulation of domestic policies to encourage capital inflows or outflows) the introduction of greater exchange rate variability has created much greater scope for economizing on reserves. The extent to which such scope has been used, has however, varied among countries and over time. As a result, economists have had much less success in estimating stable demand functions for international

reserves than in finding stable money-demand functions in individual economies.

Concerning reserve-supply arrangements, there have been two alternative views of the mechanism at work, partly reflecting changing institutional conditions in the world economy. The traditional view, which was widely accepted until the latter 1960s, was that the stock of international liquidity was to a significant extent exogenously determined. There was little dispute that this was true of gold, where the price was fixed and the available physical quantity was being augmented at a relatively slow and predictable rate. It was also thought to be broadly characteristic of US dollars, with the payments deficit of the United States supplying foreign exchange which, under the then existing institutional arrangements, other countries felt obliged to accept and hold. The newer view of reserve supply arrangements sees the stock of international liquidity as being essentially demand-determined. With international capital markets having greatly expanded in size and efficiency, countries may collectively increase their stocks of owned liquidity through operations in domestic and international financial markets. In this view, reserve stocks can increase quite independently of the balance of payments position of reserve currency countries. If a country wishes to increase its reserve holdings, and is creditworthy, it may bid for the desired funds in international financial markets and hold them in the form of short-term securities. The liabilities which are the counterpart to the reserves thus created may be liabilities of the public or private sector, and may be issued by residents of any country with the creditworthiness (and exchange control permission) to do so.

These different views of reserve-supply arrangements, coupled with differences of opinion about the stability of the underlying demand for reserves, have major implications for the role of international liquidity in influencing developments in the world economy. At one extreme, if the demand for reserves was a stable function of variables that were closely linked to world output and trade, and if the available stock of reserves was externally fixed, there would be a tight linkage between international liquidity and economic activity. If reserves fell below the desired stock for a given level of economic activity, countries would, on average, seek to augment their liquidity through trade restrictions, exchange rate depreciation, or other measures aimed at strengthening their overall balance of payments. With a fixed stock of reserves, a deflationary bias would be imparted to the world economy until nominal incomes had been reduced sufficiently to correspond with the given reserve stock.

While it was never believed that such a tight linkage existed, the fear of reserve inadequacy became important during the 1960s. Gold was fixed in price and increasing only slowly in volume, while dollars were thought to be created by a process that would eventually prove unsustainable because of the effects of continued US deficits on countries' willingness to hold dollars. For this reason, it was felt that some other mechanism was required to meet the growth in demand for reserves. Initially, devices were employed to augment the supply of credit facilities (central bank 'swap' arrangements, General Arrangements to Borrow, increased drawing rights in the IMF) but it was generally felt that the system required a secular increase in owned reserves as well. The outcome of this debate was the decision (reached in 1967) to create a new international asset, Special Drawing Rights in the International Monetary Fund (SDRs). Since SDRs would not be the liabilities of any individual country, and would have a value linked to gold (later, to a basket of currencies), they would not be subject to the confidence factors that affected US dollars. At the same time, the volume of SDRs could be augmented (or reduced) by conscious decision in the light of the long-term needs of the world economy. At the time of writing, SDRs are valued in terms of a basket of the five major industrial country currencies and bear an interest rate (paid by countries whose holdings are less than their allocation and received by those countries whose holdings exceed their allocation) related to short-term market interest rates.

Since the shift to floating exchange rates, and partly as a result of it, a somewhat different view has developed of the factors influencing the demand for and supply of reserves. It has been

recognized that the demand for reserves by major countries with flexible exchange rates cannot be easily identified, and may change over time. At the same time, the flexibility with which international capital markets have functioned has permitted changing reserve demand to be accommodated relatively easily. For these reasons, several major industrial countries have not thought it necessary for international liquidity to be deliberately augmented through allocations of SDRs.

On the other hand, most developing countries have continued to pursue some form of pegging arrangement, and many of them have experienced difficulty in preserving access to international capital markets. Their need for reserves has tended to increase in line with the volume of their international transactions, and their means of satisfying this need has relied heavily on action to improve the current account of their balance of payments. Many observers have therefore advocated continued creation of SDRs as a means of satisfying the reserve needs of these countries without requiring excessive adjustment on their part. At the same time, it has been pointed out that SDRs would preserve the 'seignorage' associated with reserve issuance for the international community at large. Proposals to 'link' liquidity creation to development assistance, by allocating SDRs in the first instance to developing countries or to development finance institutions, have enjoyed considerable popularity in the economic literature, but have not had the support of major countries.

See Also

- ▶ External Debt
- ▶ Gold Standard
- ▶ International Monetary Institutions
- ▶ International Monetary Policy

Bibliography

Crockett, A.D. 1978. *Control over international reserves*. Washington, DC: IMF Staff Papers, March.

Heller, H.R., and M.S. Khan. 1978. *The demand for international reserves under fixed and floating exchange rates*. Washington, DC: IMF Staff Papers, December.

International Monetary Fund. 1970. *International reserves: Needs and availability*. Washington, DC: International Monetary Fund.

Mundell, R.A., and J.J. Polak (eds.). 1977. *The new international monetary system*. New York: Columbia University Press.

Von Furstenberg, G.M. (ed.). 1983. *International money and credit: The policy roles*. Washington, DC: International Monetary Fund.

Willett, T.D. 1980. *International liquidity issues*. Washington, DC: American Enterprise Institute.

Williamson, J. 1973. International liquidity: A survey. *Economic Journal* 83(September): 685–746.

International Migration

George J. Borjas

Abstract

The resurgence of large-scale immigration in many countries has stimulated a great deal of research on many aspects of the economics of immigration. A key insight of economic theory is that the impact of immigration depends on how the skills of immigrants compare with those of natives in the host country. This article examines the ideas and models that are typically used to analyse flows of persons across countries, and illustrates how this framework increased our understanding of the determinants of the direction, size, and skill composition of immigrant flows, and of the consequences of those flows on economic outcomes.

Keywords

Elasticity of complementarity; Elasticity of substitution; Immigrant self-selection; Immigrant skills; Immigration and the welfare state; Immigration surplus; International migration; Labour flows; Labour markets; Migration costs; National Academy of Sciences (US); Redistribution of income; Roy model; Social insurance

JEL Classifications

J10

There was a significant resurgence in international migration in the last three decades of the twentieth century. By the end of the century, about 175 million persons – or almost three per cent of the world's population – resided in a country where they were not born. Nearly 9 per cent of the population in Germany, 11 per cent in France or Sweden, 12 per cent in the United States, 19 per cent in Canada, 23 per cent in New Zealand, and 25 per cent in Switzerland is foreign-born (United Nations 2002). These sizable labour flows altered economic opportunities for workers in both sending and receiving countries, and generated a great deal of debate over the economic impact of immigration and over the types of immigration policies that host countries should pursue.

Labour flows across labour markets –whether within or across countries – play a central role in any discussion of labour market equilibrium. These labour flows help markets attain a more efficient allocation of resources. This article surveys the economic analysis of immigration. In particular, it investigates the determinants of the immigration decision and the impact of that decision on economic conditions in the receiving country.

The discussion emphasizes the ideas and models that are used to analyse flows of persons across countries, and examines the implications of these models for empirical research and for our understanding of the labour market effects of immigration. A key insight of economic theory is that the economic impact of immigration depends on how the skills of immigrants compare with those of natives in the host country. As a result, much of the research effort in the immigration literature has been devoted to: (a) understanding the factors that determine the relative skills of the immigrant flow; (b) measuring the relative skills of immigrants in the host country; and (c) evaluating how the skill composition of the immigrant influx affects economic outcomes.

Because the discussion focuses on the impact of immigration on economic conditions in the host country, the analysis ignores a number of equally important issues, in terms of both their theoretical implications and their empirical significance. Immigration, after all, alters economic opportunities not only in the host country, but in the source country as well. Few studies, however, investigate what happens to economic opportunities in a source country when a selected subsample of its population moves elsewhere. Similarly, the discussion focuses on the economic impact of immigrants, and ignores the long-run impact of the children and grandchildren of immigrants on the host country.

The Impact of Immigration on a Host Country's Labour Market

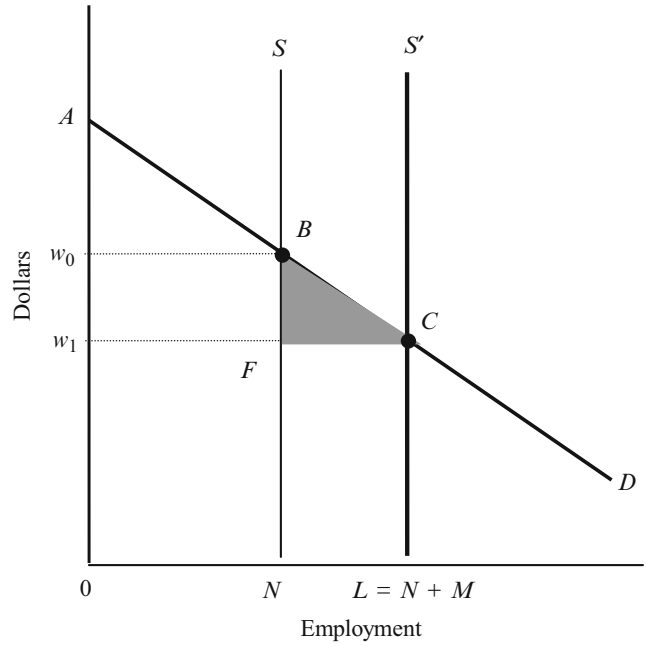
Consider initially the simplest theoretical framework that can be used to understand how immigration alters the economic rewards accruing to various factors of production in a host country. Suppose the linear homogeneous aggregate production function in the country is given by $Q = f(K, L)$, where Q is output, K is capital, and L is labour. The workforce contains N native and M immigrant workers, and all workers are perfect substitutes in production $(L = N + M)$. Natives own the entire capital stock in the host country and initially the supply of capital is perfectly inelastic. The supplies of both natives and immigrants are also perfectly inelastic. Finally, let the price of the output be the numeraire.

In a competitive equilibrium, each factor price equals the respective value of marginal product. The rental rate of capital in the pre-immigration equilibrium is $r_0 = f_K(K,\ N)$ and the price of labour is $w_0 = f_L(K,\ N)$. In the pre-immigration regime, national income accruing to natives, Q_N, is:

$$Q_N = r_0 K + w_0 N. \tag{1}$$

Figure 1 illustrates this initial equilibrium. Because the supply of capital is fixed, the area under the marginal product of labour curve (f_L) gives the value of the economy's total output. The national income accruing to natives Q_N is given by the trapezoid ABN0.

The entry of M immigrants shifts the supply curve and lowers the market wage to w_1. The area in the trapezoid ACL0 now gives national income.



International Migration, Fig. 1 The immigration surplus

Immigrants receive part of the increase in national income as labour earnings (w_1M). The area in the triangle BCF gives the increase in national income that accrues to natives, or the immigration surplus'. If we use the approximation that $(w_0 - w_1) \approx (\partial w/\partial L) \times M$, the immigration surplus as a fraction of national income equals:

$$\frac{\Delta Q_N}{Q} = -\frac{1}{2}\alpha_L \ \varepsilon_{LL} \ m^2, \tag{2}$$

where α_L is labour's share of national income $(\alpha_L = wL/Q)$; ε_{LL} is the elasticity of factor price for labour ($\varepsilon_{LL} = d \ log \ w/d \ log \ L$,with marginal cost held constant); and m is the fraction of the workforce that is foreign born $(m = M/L)$.

Equation (2) can be used to calculate how much a host country gains from immigration. In the United States, for example, the share of labour income is about 70 per cent, and the fraction of immigrants in the workforce was 13 per cent in 2000. Hamermesh's survey (1993, pp. 26–9) of the empirical evidence on labour demand suggests that the elasticity of factor price for labour may be around −.3. The US immigration surplus, therefore, is on the order of 0.2 per cent of GDP.

Immigration also redistributes income from labour to capital. As Fig. 1 shows, native workers lose the area in the rectangle w_0BFw_1, and this quantity plus the immigration surplus accrues to capitalists. The net changes in the incomes of native workers and capitalists are approximately given by:

$$\left.\frac{\text{Change in native labour earnings}}{Q}\right|_{dK=0} = \alpha_L \ \varepsilon_{LL} \ m \ (1-m), \tag{3}$$

$$\left.\frac{\text{Change in income of capitalists}}{Q}\right|_{dK=0} = -\alpha_L \ \varepsilon_{LL} m\left(1-\frac{m}{2}\right). \tag{4}$$

Consider again the back-of-an-envelope calculation for the United States. If the elasticity of factor price is −.3, native-born workers lose about 2.4 per cent of GDP, while native-owned capital gains about 2.6 per cent of GDP. A small immigration surplus may disguise a sizable income transfer from workers to the users of immigrant labour.

The derivation of the surplus in (2) assumed that the host country's capital stock is fixed. Alternatively, suppose that the supply of capital is perfectly elastic at the world price $(dr = 0)$, so that in the long run the capital stock adjusts completely to the increased labour supply. Differentiating the marginal productivity condition $r = f_K(K,L)$ implies that the immigration-induced change in the capital stock is:

$$\left.\frac{dK}{dM}\right|_{dr=0} = -\frac{f_{KL}}{f_{KK}} > 0. \tag{5}$$

The derivative in (5) is positive because $f_{KL} > 0$ when the production function is linear homogeneous.

The elasticity of complementarity for any input pair i and j is defined by $c_{ij} = f_{ij}f/f_i f_j$. (The elasticity of complementarity is the dual of the elasticity of substitution. Hamermesh 1993, Ch. 2, presents a detailed discussion of the properties of the elasticity of complementarity.) The immigration-induced wage change is then given by:

$$\left.\frac{d\log w}{d\log M}\right|_{dr=0} = \frac{\alpha_L}{c_{KK}}\left(c_{KK}\ c_{LL} - c_{LK}^2\right)m. \tag{6}$$

The linear homogeneity of the production function implies that $\left(c_{KK}\ c_{LL} - c_{LK}^2\right) = 0$, so that the host country's wage is independent of immigration. The immigration-induced capital flow re-establishes the pre-immigration capital–labour ratio in the host country. Immigration does not alter the price of labour or the returns to capital, and natives neither gain nor lose from immigration. The long-run immigration surplus is zero.

The conclusion that immigration does not alter labour market conditions in the long run depends critically on the assumption of a homogeneous labour force. Suppose there are two types of workers in the host country's labour market, skilled (L_S) and unskilled (L_U). The linear homogeneous aggregate production function is:

$$\begin{aligned} Q &= f(K,\ L_S L_U) \\ &= f[K,\ bN + \beta M,\ (1-b)N + (1-\beta)M], \end{aligned} \tag{7}$$

where b and β denote the fraction of skilled workers among natives and immigrants, respectively. The price of each factor of production, r for capital and $w_i (i = S,\ U)$ for labour, is determined by the respective marginal productivity condition. The assumption that r is fixed implies that the immigration-induced adjustment in the capital stock equals (see Borjas 1999, pp. 1703–5):

$$\left.\frac{dK}{dM}\right|_{dr=0} = -\frac{[f_{KS}\beta + f_{KU}(1-\beta)]}{f_{KK}}. \tag{8}$$

We can determine the impact of immigration on the wage of skilled and unskilled workers by differentiating the respective marginal productivity conditions and by imposing the restriction in Eq. (8). The wage effects of immigration are:

$$\begin{aligned} \frac{d\ log\ w_S}{d\ log\ M}|_{dr=0} &= \frac{\alpha_S}{c_{KK}}\left[c_{SS}c_{KK} - c_{SK}^2\right] \\ &\quad \times \frac{(\beta - b)}{p_S\, p_U}(1-m)m, \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{d\log w_U}{d\log M}|_{dr=0} &= \frac{-\alpha_U}{c_{KK}}\left[c_{UU}c_{KK} - c_{UK}^2\right] \\ &\quad \times \frac{(\beta - b)}{psp_U}(1-m)\ m, \end{aligned} \tag{10}$$

where α_i is the share of national income accruing to factor i; and p_S and p_U are the shares of the workforce that are skilled and unskilled, respectively.

The assumption that the isoquants between any pair of inputs in the production function $f(K,\ L_S,\ L_U)$ have the typical convex shape implies that $c_{11}\ c_{22} - c_{12}^2 > 0$. Equations (9) and (10) then reveal that the impact of immigration on the wage structure depends on how the skill distribution of immigrants compares with that of natives. If the two skill distributions are equal $(\beta = b)$, immigration has no impact on the wage structure of the

host country. If immigrants are relatively unskilled $(\beta < b)$, the unskilled wage declines and the skilled wage rises. If immigrants are relatively skilled $(\beta > b)$, the skilled wage declines and the unskilled wage rises. In the long run, therefore, immigration lowers the wage of substitutes and raises the wage of complements.

It can be shown that the immigration surplus as a fraction of national income is given by:

$$\frac{\Delta Q_N}{Q}|_{dr=0} = \frac{-\alpha_S^2}{2c_{KK}}\left[c_{SS}c_{KK} - c_{SK}^2\right] \times \frac{(\beta - b)^2}{p_S^2 p_U^2}(1-m)^2 m^2. \quad (11)$$

The immigration surplus is zero if $\beta = b$, and positive if $\beta \neq b$. If immigrants had the same skill distribution as natives, the immigration-induced change in the capital stock implies that the wages of skilled and unskilled workers are unaffected by immigration. The gains arise only if immigrants differ from natives.

Some studies simulate this model to provide back-of-an-envelope calculations of the immigration surplus when there is heterogeneous labour (Borjas 1995; Johnson 1997). In the US context, the immigration surplus calculated in this more general setting is roughly of the same order of magnitude (less than 0.2 per cent of GDP) as that estimated from the simplest framework illustrated in Fig. 1. The available evidence, therefore, suggests that the net measurable gains from immigration to the United States tend to be small.

Finally, it is worth emphasizing that any credible estimate of the economic benefits from immigration must rely on a theoretical framework that fully captures the various effects which inevitably arise as the impact of immigration ripples through the economy. Inevitably, different models of the economy will lead to different estimates of the economic benefits. Recent theoretical work by trade economists, for example, suggests that if one takes the Ricardian perspective that the United States provides superior economic opportunities for all factors of production – so that both capital and labour would get higher returns by migrating to the United States – immigration would actually *lower* the GDP accruing to natives substantially, by around 1.0 per cent of GDP (Davis and Weinstein 2002). Therefore, the important point to draw from the existing evidence is that plausible models of the US economy indicate that, at best, the net gains from immigration for the native-born population are very small.

Estimating the Labour Market Impact of Immigration

As shown above, economic theory suggests that immigration into a particular labour market affects the wage structure by raising the wage of complementary workers and lowering the wage of substitutes. Almost all of the first-generation empirical studies in the literature define the labour market along a geographic dimension, such as metropolitan areas in the United States. Beginning with Grossman (1982), the typical study regresses a measure of native economic outcomes in the locality (or the change in that outcome) on the relative quantity of immigrants in that locality (or the change in the relative number). (Representative studies include Altonji and Card 1991; Card 1990; 2001; Pischke and Velling 1997.) The regression coefficient is then interpreted as the impact of immigration on the native wage structure.

This approach has two well-known problems. First, immigrants may not be randomly distributed across labour markets. If immigrants endogenously cluster in areas that have done well over some time periods, this would produce a positive spurious correlation between immigration and area outcomes either in the cross-section or in the time series. Second, natives may respond to the entry of immigrants in a local labour market by moving their labour or capital to other localities until native wages and returns to capital are again equalized across areas. For example, a large immigrant flow arriving in California might well result in fewer workers moving to California, as well as a reallocation of capital from other states into California. Interregional comparisons of the wage of native workers might show little or no

difference because the effects of immigration are diffused throughout the national economy.

In view of these potential problems it is not too surprising that the region-based empirical literature has produced a confusing array of results (see the survey in Friedberg and Hunt 1995). Nevertheless, there is a tendency for the estimated cross- region correlations to cluster around zero, creating the conventional wisdom that immigrants have little impact on the labour market opportunities of native workers. It would seem, therefore, that a fundamental implication of the competitive model of the labour market – that supply shocks alter the wage structure – is soundly rejected by the data.

Because local labour markets adjust to immigration, recent research emphasizes that the labour market impact of immigration may be measurable only at the national level. Borjas (2003) used this insight to derive an estimable framework that can be used to measure the national labour market effects of immigration by linking the evolution of the wage structure in the host country to changes in immigration. As an illustration, suppose that the national workforce in the host country is composed of skill groups defined in terms of both educational attainment and work experience. The aggregate production function at time t is:

$$Q_t = \left[\lambda_{Kt} K_t^{v} + \lambda_{Lt} L_t^{v}\right]^{1/v}, \tag{12}$$

where Q is output, K is capital, L denotes the aggregate labour input; and $v = 1 - 1/\sigma_{KL}$, with σ_{KL} being the elasticity of substitution between capital and labour $(-\infty < v \leq 1)$. The vector λ gives technology parameters that shift the production frontier, with $\lambda_{Kt} + \lambda_{Lt} = 1$. The aggregate L_t incorporates the contributions of workers who differ in both education and experience. Let:

$$L_t = \left[\sum_i \theta_{it} L_{it}^{\rho}\right]^{1/\rho} \tag{13}$$

where L_{it} gives the number of workers with education i at time t, and $\rho = 1 - 1/\sigma_E$, with σ_E being the elasticity of substitution across these education aggregates $(-\infty < \rho \leq 1)$. The θ_{it} give time-variant technology parameters that shift the relative productivity of education groups, with $\Sigma_j \theta_{it} = 1$. Finally, the supply of workers in each education group is itself given by an aggregation of the contribution of similarly educated workers with different experience. In particular,

$$L_{it} = \left[\sum_j \alpha_{ij} L_{ijt}^{\eta}\right]^{1/\eta}, \tag{14}$$

where L_{ijt} gives the number of workers in education group i and experience group j at time t (given by the sum of N_{ijt} native and M_{ijt} immigrant workers); and $\eta = 1 - 1/\sigma_X$, with σ_X being the elasticity of substitution across experience classes within an education group $(-\infty < \eta \leq 1)$. Equation (14) assumes that the technology coefficients α_{ij} are constant over time, with $\Sigma_j \alpha_{ij} = 1$.

The three elasticities of substitution that summarize all the economically relevant information in the production technology can be easily estimated using data on factor prices and quantities. The empirical application of this framework to US Census data from 1960 through 2000 in Borjas (2003) indicated that $\sigma_X = 3.5$, $\sigma_E = 1.3$, and $\sigma_{KL} = 1.0$. These elasticity estimates, combined with estimates of the size of the immigrant influx for each skill group, can be used to calculate the impact of immigration on the wage structure in a host country. Define the factor price elasticity giving the impact on the wage of factor y of an increase in the supply of factor z as:

$$\varepsilon_{yz} = \frac{d \log w_y}{d \log L_Z}, \tag{15}$$

It is easy to show that the factor price elasticities depend on the income shares accruing to the various factors and on the three elasticities of substitution in the three-level constant elasticity of substitution (CES) framework. The marginal productivity condition for the typical worker in education group s and experience group x can be written as $w_{sx} = D(K, L_{ij})$, where L_{ij} is a vector indicating the number of workers in each of the education–experience cells. Suppose that the

capital stock is constant. The short-run impact of immigration on the log wage of group $(s,\ x)$ is:

$$\Delta \log w_{sx} = \sum_i \sum_j \varepsilon_{sx,ij}\ m_{ij}, \qquad (16)$$

where m_{ij} gives the percentage change in labour supply due to immigration in skill cell (i, j). The available evidence suggests that the 1980–2000 immigrant influx into the United States, which represented an 11 per cent increase in labour supply, lowered the wage of the typical native worker by 3.7 per cent in the short run. As indicated in the earlier discussion, the adverse wage effects of immigration on the average worker are muted as the capital stock adjusts to the supply shock.

The Self-selection of Immigrants

As we have seen, the economic impact of immigration depends crucially on the differences in the skill distributions of immigrants and natives. Not surprisingly, a great deal of research effort has focused on the question of how immigrant skills compare with those of native workers. Perhaps the central finding of this literature is that immigrants are not a randomly selected sample of the population of the source countries.

It is instructive to consider a two-country model (Borjas 1987). Residents of the source country (country 0) consider migrating to the host country (country 1). Assume the migration decision to be irreversible. Residents of the source country face the earnings distribution:

$$\log w_0 = \mu_0 + v_0, \qquad (17)$$

where w_0 gives the wage in the source country; μ_0 gives the mean earnings in the source country; and the random variable v_0 measures deviations from mean earnings and is normally distributed with mean zero and variance σ_0^2. For simplicity, Eq. (17) omits the subscript that indexes a particular individual.

Suppose the potential earnings in the host country of emigrants from country 0 can be represented by:

$$\log w_1 = \mu_1 + v_1, \qquad (18)$$

where μ_1 gives the mean earnings in the host country *for this particular population*, and the random variable v_1 is normally distributed with mean zero and variance σ_1^2. The correlation coefficient between v_0 and v_1 equals ρ_{01}.

The mean μ_1 does not necessarily equal the mean earnings of native workers in the host country. After all, the average worker in the source country might be more or less skilled than the average worker in the host country. It is convenient to assume that the average person in both countries is equally skilled (or, equivalently, that any differences in average skills have been controlled for), so that μ_1 also gives the mean earnings of natives in the host country. This assumption helps isolate the impact of the selection process on the skill composition of the immigrant influx.

Equations (17) and (18) describe the earnings opportunities available to persons born in the source country. Assume that the migration decision is determined by a comparison of earnings opportunities across countries, net of migration costs. Define the index function:

$$I = \log\left(\frac{w_1}{w_0 + C}\right) \approx (\mu_1 - \mu_0 - \pi) + (v_1 - v_0), \qquad (19)$$

where C gives migration costs, and π gives a 'time-equivalent' measure of these costs $(\pi = C/w_0)$. A person emigrates if $I > 0$, and remains in the source country otherwise.

Migration is costly. Because the costs vary greatly among persons, and include direct costs, forgone earnings, and psychic costs, the sign of the correlation between costs and wages is unknowable. The distribution of the random variable π in the source country's population is:

$$\pi = \mu_\pi + v_\pi, \qquad (20)$$

where μ_π is the mean level of migration costs in the population, and v_π is a normally distributed random variable with mean zero and variance

σ_π^2. The correlation coefficients between v_π and $(v_0,\ v_1)$ are given by $(\rho_{\pi 0},\ \rho_{\pi 1})$. The probability that a person migrates to the host country can be written as:

$$P(z) = Pr[v > -(\mu_1 - \mu_0 - \mu_\pi)] = 1 - \Phi(z), \tag{21}$$

where $v = v_1 - v_0 - v_\pi,\ z = -(\mu_1 - \mu_0 - \mu_\pi)/\sigma_v$, and Φ is the standard normal distribution function.

It is easy to show that the emigration rate falls when the mean income in the source country rises, when the mean income in the host country falls, and when time-equivalent migration costs rise. Most studies in the literature on the internal migration of persons within a particular country focus on testing these theoretical predictions. The empirical evidence in these studies is generally supportive of the theory.

Although it is important to determine the size and direction of migration flows, it is equally important to determine *which* persons find it most worthwhile to migrate to the host country. This question lies at the heart of the Roy model (Roy 1951). Consider the conditional means $E(\log w_0|\mu_0,\ I > 0)$ and $E(\log w_1|\mu_1,\ I > 0)$. These means give the average earnings in both the source and host countries for persons who migrate. Note that the conditional means hold μ_0 and μ_1 constant. The calculation effectively assumes that the migration flow is sufficiently small so that there are no feedback effects on the performance of immigrants (or natives) in the host country or on the performance of the 'stayers' in the source country. Because the random variables $v_0,\ v_1$, and v_π are jointly normally distributed, these conditional means are given by:

$$E(\log w_0|\mu_0, I > 0) = \mu_0 + \left[\frac{\sigma_0\sigma_1}{\sigma_v}\left(\rho_{01} - \frac{\sigma_0}{\sigma_1}\right) - \rho_{\pi 0}\frac{\sigma_\pi}{\sigma_1}\right]\lambda, \tag{22}$$

$$E(\log w_1|\mu_1, I > 0) = \mu_1 + \left[\frac{\sigma_0\sigma_1}{\sigma_v}\left(\frac{\sigma_1}{\sigma_0} - \rho_{01}\right) - \rho_{\pi 1}\frac{\sigma_\pi}{\sigma_0}\right]\lambda, \tag{23}$$

where $\lambda = \varphi(z)/(1 - \Phi(z))$, and φ is the density of the standard normal. It is easier to interpret the results in Eqs. (22) and (23) by assuming that $\sigma_\pi = 0$, so that time-equivalent migration costs are constant. Let $Q_0 = E(v_0|\mu_0,\ I > 0)$ and $Q_1 = E(v_1|\mu_1,\ I > 0)$. The Roy model identifies three cases that summarize the skill differentials between immigrants and natives:

$$\begin{aligned} &Q_0 > 0 \text{ and } Q_1 > 0, \quad \text{if} \quad \rho_{01} > \frac{\sigma_0}{\sigma_1} \text{ and } \frac{\sigma_1}{\sigma_0} > 1,\\ &Q_0 < 0 \text{ and } Q_1 < 0, \quad \text{if} \quad \rho_{01} > \frac{\sigma_0}{\sigma_1} \text{ and } \frac{\sigma_1}{\sigma_0} > 1,\\ &Q_0 < 0 \text{ and } Q_1 > 0, \quad \text{if} \quad \rho_{01} < \min\left(\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1}\right). \end{aligned} \tag{24}$$

Positive selection occurs when immigrants have above-average earnings in both the source and host countries ($Q_0 > 0$ and $Q_1 > 0$), and negative selection when immigrants have below-average earnings in both countries ($Q_0 < 0$ and $Q_1 < 0$). Equation (24) shows that either type of selection requires that skills be positively correlated across countries. The standard deviations σ_0 and σ_1 measure the 'price' of skills: the greater the rewards to skills, the larger the inequality in wages. Immigrants are then positively selected when the source country – *relative to the host country* – 'taxes' highly skilled workers and 'insures' less skilled workers from poor labour market outcomes, and immigrants are negatively selected when the host country taxes highly skilled workers and subsidizes less skilled workers.

There also exists the possibility that the host country draws persons who have below-average earnings in the source country but do well in the host country ($Q_0 < 0$ and $Q_1 > 0$). This sorting occurs when the correlation coefficient ρ_{01} is small or negative. This correlation may be negative when a source country experiences a structural political shift, such as a Communist takeover. In its initial stages, this political system often redistributes incomes by confiscating the assets of relatively successful persons. Immigrants from such systems will be in the lower tail of the post-revolution income distribution, but will perform well in a host country's market economy.

Equation (24) shows that neither differences in mean incomes across countries nor the level of migration costs determine the type of selection that characterizes immigrants. Mean incomes and migration costs affect the size of the flow (and the extent to which the skills of the average immigrant differ from the mean skills of the population), but they do not determine whether the immigrants are drawn mainly from the upper or lower tail of the skill distribution.

The discussion assumed that migration costs are constant in the population. Variable migration costs do not alter any of the selection rules if (a) time-equivalent migration costs are uncorrelated with skills, or (b) the variation in migration costs is 'small' relative to the variation in earnings. Otherwise, variable migration costs can change the nature of selection. Suppose that τ is negatively correlated with earnings, perhaps because less skilled persons find it more difficult to find jobs in the host country. This negative correlation increases the likelihood that the immigrant flow is positively selected.

Some of the implications of the Roy model have been tested empirically by estimating the correlation between the earnings of immigrants in a host country, typically the United States, and measures of the rate of return to skills in source countries. The evidence provides mixed support for the Roy model's prediction that immigrants originating in countries with higher rates of return to skills have lower earnings in the United States. Borjas (1987) reports that measures of income inequality in the source country, which are a very rough proxy for the rate of return to skills, are weakly negatively correlated with the earnings of immigrant men, while Cobb-Clark (1993) reports a similar finding for immigrant women. In contrast, Chiquiar and Hanson's study (2005) of Mexican emigration finds that the least skilled persons in the Mexican population tend to be under-represented in the Mexican-born workforce in the United States. Because the Mexican wage distribution has a larger variance than that of the United States, these low-skill workers would presumably be the persons most motivated to emigrate. This finding suggests that non-constant migration costs in the population of some source countries may play an important role in determining the selection of immigrants.

Measuring Trends in Immigrant Skills

Beginning with the work of Chiswick (1978), a large literature has developed that attempts to measure the skill differential between immigrants and natives at the time of entry and how this differential changes over time as immigrants adapt to the host country's labour market. A key result of this literature is that there exists a cross-sectional positive correlation between the earnings of immigrants and the number of years that have elapsed since immigration. As will be seen below, there has been a great deal of debate over the interpretation of this correlation.

The empirical analysis of the relative economic performance of immigrants was initially based on the cross-section regression model:

$$\log w_\ell = X_\ell \beta_0 + \beta_1 I_\ell + \beta_2 y_\ell + \in \ell, \qquad (25)$$

where w_ℓ is the wage rate of person ℓ in the host country; X_ℓ is a vector of socioeconomic characteristics (which includes age); I_ℓ is a dummy variable set to unity if person ℓ is foreign-born; and y_ℓ gives the number of years that the immigrant has resided in the United States and is set to zero if ℓ is a native. (The models used in empirical studies typically include higher-order polynomials in age and years-since-migration. These nonlinearities, however, do not play a role in the identification issue discussed below.) Because the vector X includes the worker's age, the coefficient β_2 measures the differential value that the host country's labour market attaches to time spent in the host country versus time spent in the source country.

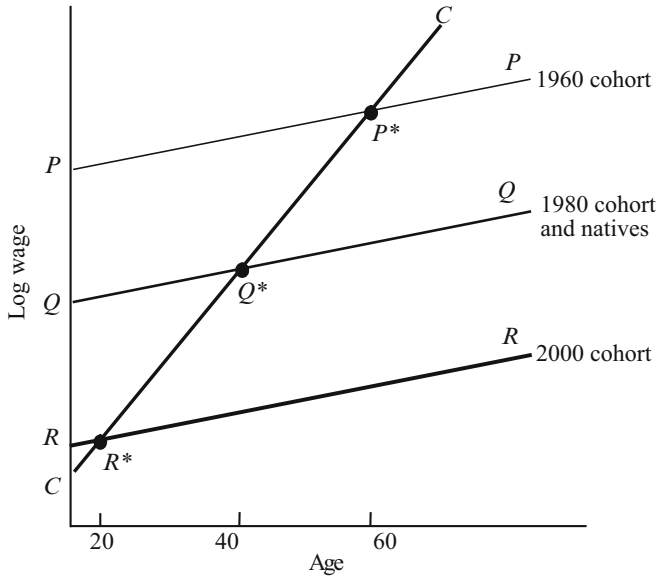
Cross-section studies of immigrant earnings in several host countries have typically found that β_1 is negative and β_2 is positive. (Although most of the empirical evidence focuses on the US experience, the literature finds a similar correlation in Canada – Baker and Benjamin 1994; Australia – Beggs and Chapman 1991; and Germany – Dustmann 1993.) Chiswick's (1978) analysis of

the 1970 US Census data indicates that immigrants earn about 17 per cent less than 'comparable' natives at the time of entry, and this gap narrows by slightly over one percentage point per year. As a result, immigrant earnings overtake those of their native counterparts after about 15 years in the United States. The steeper age-earnings profiles of immigrants was interpreted as meaning that immigrants accumulated more human capital than natives as the assimilation process took hold, closing the wage gap between the two groups. The overtaking phenomenon was then explained by assuming that immigrants were positively selected. As we have seen, this assumption about the selection process is not necessarily implied by income-maximizing behaviour on the part of immigrants.

Borjas (1985) suggested an alternative interpretation of the cross-section evidence. Instead of interpreting the positive β_2 as a measure of assimilation, he argued that the cross-section data might be revealing a decline in relative skills across successive immigrant cohorts. In the United States, the post-war era witnessed major changes in immigration policy and in the size and national origin mix of the immigrant flow. If these changes generated a less-skilled immigrant flow, the cross- section correlation indicating that more recent immigrants earn less may say little about the process of wage convergence, but may instead reflect innate differences in ability or skills across cohorts.

To illustrate the identification problem, consider a hypothetical situation where there are three separate immigrant waves, and these waves have distinct productivities. One wave arrived in 1960, the second arrived in 1980, and the last arrived in 2000. Suppose also that all immigrants enter the United States at age 20.

Assume that the earliest cohort has the highest productivity level of any group in the population, including US-born workers. If we could observe their earnings in every year after they arrive in the United States, their age-earnings profile would be given by the line *PP* in Fig. 2. For the sake of argument, let's assume that the last wave of immigrants (that is, the 2000 arrivals) is the least productive of any group in the population, including natives. If we could observe their earnings throughout their working lives, their age-earnings profile would be given by the line *RR* in the figure. Finally, suppose that the immigrants who arrived in 1980 have the same skills as natives. If we could observe their earnings at every age in their working lives, the age-earnings profiles of this cohort and of natives would overlap and be given by the line *QQ*. Note that the age-earnings profiles of each of the immigrant cohorts is



International Migration, Fig. 2 Cohort effects and the immigrant age-earnings profile

parallel to the age-earnings profile of the native population. There is no wage convergence between immigrants and natives in this hypothetical example.

Suppose we now have access to data drawn from the 2000 decennial census. This cross-section data-set, which provides a snapshot of the US workforce as of 1 April 2000, provides information on each worker's wage rate, age, whether native- or foreign-born, and on the year the worker arrived in the United States. As a result, we can observe the wage of immigrants who have just arrived as part of the 2000 cohort when they are 20 years old (see point R^* in Fig. 2). We can also observe the wage of immigrants who arrived in 1980 when they are 40 years old (point Q^*), and we observe the wage of immigrants who arrived in 1960 when they are 60 years old (point P^*). A cross-section data-set, therefore, allows us to observe only one point on each of the immigrant age-earnings profiles.

If we connect points P^*, Q^*, and R^*, we trace out the immigrant age-earnings profile that is generated by the cross-sectional data, or line CC in Fig. 2. This cross-section line has two important properties. First, it is substantially steeper than the native age-earnings profile. The tracing out of the age-earnings profile of immigrants using cross-section data makes it seem as if there is wage convergence between immigrants and natives, when in fact there is none. Second, the cross-section line CC crosses the native line at age 40. This gives the appearance that immigrant earnings overtake those of natives after they have been in the United States for 20 years. In fact, no immigrant group experienced such an overtaking.

The identification of aging and cohort effects raises difficult methodological problems in many demographic contexts. Identification requires the availability of longitudinal data where a particular worker is tracked over time, or, equivalently, the availability of a number of repeated cross-sections so that specific cohorts can be tracked across survey years. Suppose that a total of Ω cross-section surveys are available, with cross-section $\tau(\tau = 1, \ldots, \Omega)$ being obtained in calendar year T_τ. Pool the data for immigrants and natives across the cross-sections, and consider the regression model:

$$\begin{aligned}\text{Immigrant equation}: \log w_{\ell\tau} &= X_{l\tau}\varphi_{i\tau} + \alpha y_{\ell\tau} \\ &\quad + \beta C_{\ell\tau} \\ &\quad + \sum_{\tau=1}^{\Omega} \gamma_{i\tau}\pi_{\ell\tau} + \varepsilon_{\ell\tau}, \quad (26)\end{aligned}$$

$$\begin{aligned}\text{Native equation}: \log w_{\ell\tau} &= X_{\ell\tau}\varphi_{n\tau} + \sum_{\tau=1}^{\Omega}\gamma_{n\tau}\ \pi_{\ell\tau} \\ &\quad + \varepsilon_{\ell\tau}, \quad (27)\end{aligned}$$

where $w_{\ell\tau}$ gives the wage of person ℓ in cross-section τ, X gives a vector of socio-economic characteristics (including age); $C_{\ell\tau}$ gives the calendar year in which the immigrant arrived in the host country; $y_{\ell\tau}$ gives the number of years that the immigrant has resided in the host country $(y_{\ell\tau} = T_\tau - C_{\ell\tau})$; and $\pi_{\ell\tau}$ is a dummy variable indicating if person ℓ was drawn from cross-section τ.

The identification problem arises from the identity:

$$y_{l\tau} \equiv \sum_{\tau=1}^{\Omega}\ \pi_\tau(T_\tau - C_{\ell\tau}). \quad (28)$$

Equation (28) introduces perfect collinearity among the variables $y_{\ell\tau}$, $C_{\ell\tau}$ and $\pi_{\ell\tau}$ in the immigrant earnings function. As a result, the key parameters of interest – α, β, and the vector γ_j – are not identified. Some type of restriction must be imposed to separately identify the aging effect, the cohort effect, and the period effects. Borjas (1985) proposed the restriction that the period effects are the same for immigrants and natives:

$$\gamma_{i\tau} = \gamma_{n\tau}, \qquad \forall\tau. \quad (29)$$

Put differently, trends in aggregate economic conditions change immigrant and native wages by the same percentage amount. A useful way of

thinking about this restriction is that the period effects for immigrants are calculated from *outside* the immigrant wage determination system.

The measurement of cohort and assimilation effects has received a great deal of attention, particularly in the US, context where the data indicate that cross-section age-earnings profiles overestimate the rate of convergence between immigrant and native earnings due to the presence of cohort effects like those illustrated in Fig. 2.

Immigration and the Welfare State

In addition to the labour market consequences, immigration has fiscal impacts on host countries because there may be significant costs associated with providing social services to the immigrants, and these costs will depend both on the skill composition of the immigrant population and on the generosity of the host country's welfare state. In fact, the immigration debate in many receiving countries has often focused on the possibility that immigrants may become public charges. Since 1882, for example, the United States has banned the entry of 'any persons unable to take care of himself or herself without becoming a public charge'. Similarly, the US immigration statutes declare that 'any alien who, within five years after the date of entry, has become a public charge . . . is deportable'.

There has been a great deal of concern over the possibility that the relatively generous welfare programmes offered by industrialized Western economies have become a magnet for immigrants. It is possible, for example, that generous welfare programmes attract immigrants who otherwise would not have migrated, or that the safety net discourages immigrants who 'fail' in the host country from returning to their origin. These magnetic effects raise questions about both the political legitimacy and economic viability of the welfare state. Who is entitled to the safety net that the host country's taxpayers pay for? And can the richer host countries afford to extend that safety net to the population of poorer countries? Surprisingly, few studies attempt to determine whether such magnetic effects are empirically important.

Much of the empirical debate over the link between immigration and welfare in recent years has instead been dominated by the bottom line: do immigrants pay their way in the welfare state? There exist a large number of accounting exercises, each purporting to calculate the amount of taxes paid by immigrants and the amount of social expenditures that can be attributed to immigrants. The estimates provided by many of these studies are often unconvincing, with the conclusion typically dictated by the accounting assumptions employed in the exercise. For example, how does one allocate expenditures in various public goods between immigrants and natives? In 1996, the US National Academy of Sciences attempted to settle the issue by examining in detail this contentious issue (Smith and Edmonston 1997, Ch. 6).

The National Academy report measured the 'short-run' impact of immigration on the fiscal ledger sheet of states and local governments, that is, the fiscal impact during a particular fiscal year. For two major immigrant-receiving states, California and New Jersey, the National Academy conducted an item-by-item accounting of expenditures incurred and taxes collected, and calculated how immigration affected each of these entries.

California attracts a disproportionately large number of the welfare recipients in the immigrant population, and provides a wide array of expensive services, ranging from generous welfare assistance to a world-class system of public universities and a sophisticated and well-maintained system of roads and freeways. It turns out that immigration increased the state and local taxes paid by the typical native household in California in 1995 by almost $1,174 annually. The cost-benefit calculation for New Jersey is less dramatic. Because New Jersey provides fewer state and local services, and because New Jersey attracts a different type of immigrant (more skilled and less prone to use government services), immigration increased the annual tax bill of New Jersey's typical native household by only $229.

If one were to extrapolate these estimates nationwide, the National Academy concluded

that immigration increased the taxes of the typical native household in the United States by around $200 annually in the mid-1990s. There are approximately 90 million native households in the United States, so that the national fiscal burden is around $18 billion per year. Recall that the annual immigration surplus in the United States is estimated to be around 0.2 per cent of GDP, or roughly around $20 billion in 2000. In the short run, therefore, the available evidence suggests that the net gain (to the native population) from immigration is essentially zero.

It is important to note that this type of accounting exercise is myopic because it does not consider the long-run impact of immigration on government expenditures. For example, it has been argued that immigration may provide an important mechanism to alleviate the fiscal crisis that most industrialized countries will face as their populations age and the dependency ratio rises, putting much greater pressures on social insurance and the fiscal solvency of the welfare state. However, careful simulations of the fiscal consequences of this demographic transition –and of the costs and benefits from immigration in industrialized economies –suggest that immigration can play only a limited role in alleviating the fiscal stress.

Using an overlapping generations framework, these studies examine how the payroll tax rate must adjust to cover the expenses that will be inevitably incurred over the twenty-first century to provide social benefits to a relatively larger aging population (Fehr et al. 2004; Storesletten 2000). One can then simulate the impact of different immigration scenarios on the required payroll tax rate. These simulations typically suggest that the social insurance tax rate in industrialized economies will not fall drastically even if immigration were greatly expanded (such as doubling the size of the flow) over the next century.

The reason for the relative unimportance of immigration in this fiscal exercise is that immigrants themselves generate an increase in social expenditures, and this increase may reduce much of the perceived benefit from simply having a larger population over which to amortize the required expenses. In addition, social insurance programmes in many industrialized host countries tend to be progressive, so that the immigrant population, which is relatively low-skill, will generally contribute less to their funding and receive higher benefits. In short, immigration is not the panacea that can resolve the fiscal problems associated with an aging population in these societies.

See Also

- ▶ Economic Demography
- ▶ Globalization
- ▶ Roy Model

Bibliography

Altonji, J., and D. Card. 1991. The effects of immigration on the labor market outcomes of less-skilled natives. In *Immigration, trade, and the labor market*, ed. J. Abowd and R. Freeman. Chicago: University of Chicago Press.

Baker, M., and D. Benjamin. 1994. The performance of immigrants in the Canadian labor market. *Journal of Labor Economics* 12: 369–405.

Beggs, J., and B. Chapman. 1991. Male immigrant wage and unemployment experience in Australia. In *Immigration, trade, and the labor market*, ed. J. Abowd and R. Freeman. Chicago: University of Chicago Press.

Borjas, G. 1985. Assimilation, changes in cohort quality, and the earnings of immigrants. *Journal of Labor Economics* 3: 463–489.

Borjas, G. 1987. Self-selection and the earnings of immigrants. *American Economic Review* 77: 531–553.

Borjas, G. 1995. The economic benefits from immigration. *Journal of Economic Perspectives* 9(2): 3–22.

Borjas, G. 1999. The economic analysis of immigration. In *Handbook of labor economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: Elsevier.

Borjas, G. 2003. The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market. *Quarterly Journal of Economics* 118: 1335–1374.

Card, D. 1990. The impact of the Mariel Boatlift on the Miami labor market. *Industrial and Labor Relations Review* 43: 245–257.

Card, D. 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics* 19: 22–64.

Chiquiar, D., and G. Hanson. 2005. International migration, self-selection, and the distribution of wages: Evidence from Mexico and the United States. *Journal of Political Economy* 113: 239–281.

Chiswick, B. 1978. The effect of Americanization on the earnings of foreign-born men. *Journal of Political Economy* 86: 897–921.

Cobb-Clark, D. 1993. Immigrant selectivity and wages: The evidence for women. *American Economic Review* 83: 986–993.
Davis, D., and D. Weinstein. 2002. Technological superiority and the losses from migration. Working Paper No. 8971. Cambridge, MA: NBER.
Dustmann, C. 1993. Earnings adjustment of temporary immigrants. *Journal of Population Economics* 6: 153–168.
Fehr, H., S. Jokisch, and L. Kotlikoff. 2004. The role of immigration in dealing with the developed worlds demographic transition. *FinanzArchiv* 69: 296–324.
Friedberg, R., and J. Hunt. 1995. The impact of immigration on host county wages, employment and growth. *Journal of Economic Perspectives* 9(2): 23–44.
Grossman, J. 1982. The substitutability of natives and immigrants in production. *The Review of Economics and Statistics* 54: 596–603.
Hamermesh, D. 1993. *Labor demand*. Princeton: Princeton University Press.
Johnson, G. 1997. Estimation of the impact of immigration on the distribution of income among minorities and others. In *Help or hindrance? The economic implications of immigration for African-Americans*, ed. D. Hamermesh and F. Bean. New York: Russell Sage Press.
Pischke, J.-S., and J. Velling. 1997. Employment effects of immigration to Germany: An analysis based on local labor markets. *The Review of Economics and Statistics* 79: 594–604.
Roy, A. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3: 135–146.
Smith, J., and B. Edmonston (eds.). 1997. *The new Americans: Economic, demographic, and fiscal effects of immigration*. Washington, DC: National Academy Press.
Storesletten, K. 2000. Sustaining fiscal policy through immigration. *Journal of Political Economy* 108: 300–323.
United Nations. 2002. *International migration, 2002*. New York: Population Division, Department of Economics and Social Affairs, United Nations.

International Monetary Fund

Brett House, David Vines and W. Max Corden

Abstract

The International Monetary Fund (IMF) was set up in 1944 and charged with supervising the post-war Bretton Woods system of pegged but adjustable exchange rates as a means of promoting international monetary cooperation. Since the Bretton Woods system broke down in 1971, the IMF's role has become more complicated. It has exercised surveillance over its members' policies, worked to ensure the stability of the international financial system, and assisted the world's poorest economies. This article reviews the history and achievements of the IMF as well as the challenges it faces in redefining its role at the beginning of the 21st century.

Keywords

Balance of payments disequilibria; Balanced growth; Banking crises; Beggar-thy-neighbour; Bretton Woods system; Capital account liberalization; Capital controls; Catch-up; Clearing union; Concessionality; Conditionality; Convertibility; Currency crises; European monetary union; Exchange rate mechanism (EMU); Fixed exchange rates; Floating exchange rates; Foreign aid; General agreement on tariffs and trade; Gold exchange standard; Gold standard; Incomes policy; Inflation; Inflation targeting; International capital flows; International monetary fund; Intertemporal trade; Keynes, J.M.; Keynesianism; Meade, J.E; Monetarism; Moral hazard; NAFTA; Nominal exchange rates; Pegged exchange rates; Poverty alleviation; Price control; Real exchange rates; Sovereign debt; Special drawing rights (IMF); Taylor doctrine; Trade liberalization; World Bank

JEL Classifications

F2

The International Monetary Fund (henceforth 'the IMF' or 'the Fund') was conceived at a conference at the Mount Washington Hotel in Bretton Woods, New Hampshire, in July 1944 and its Articles of Agreement entered into force in December 1945. The World Bank (henceforth 'the Bank') was set up at the same time. The IMF was established to promote international monetary cooperation and the elimination of exchange restrictions on current account

transactions; to facilitate trade, economic growth and high levels of employment; to foster exchange rate stability; and to provide temporary financial assistance to countries so as to ease balance of payments adjustment. More specifically, it was given the role of supervising a system of pegged but adjustable exchange rates, which became known as the Bretton Woods system. In the first two sections of this entry we explain how the Bretton Woods system worked, and why it broke down in 1971. In the following sections we consider the roles which the Fund now plays, which differ from its original activities. They are: surveillance, ensuring stability for the international financial system and for individual economies within this system, and assisting the world's poorest economies. As part of each of these three activities, the Fund also provides policy advice and technical assistance. This is a much less clear collection of responsibilities, and, as a result, the future direction of the Fund is somewhat uncertain. The aim of this article is to review the achievements of the Fund, and also the challenges that lie ahead. A related overview of some of the issues discussed here can be found in Gilbert and Vines (2004).

The Bretton Woods system

Intentions

As the Second World War drew to a close, the United Kingdom, the United States and their allies, inspired in part by the *General Theory* of John Maynard Keynes (Keynes 1936), established a policy framework in which countries would be able to promote high levels of employment and output, by means of demand management policies, focused mainly on fiscal measures. This would – it was hoped – avert slumps in growth and would thereby prevent the re-emergence of the kind of global depression that had occurred in the 1930s. (See Williamson 1983a; Moggridge 1986.)

From early on, Keynes had seen that such policies would need global support. This is because they would have to be reconciled with the need for each country to be sufficiently competitive; that is, each country would need to be able to export enough to pay for the imports that would be purchased at full employment. In 1942, Keynes put forward plans for a new post-war international monetary system designed to make this possible, which he called a 'Clearing Union'. (See Keynes 1971–88, vol. 25, pp. 41–67; van Dormael 1978; Gardner 1956.) His plan drew on the theoretical arguments in his *General Theory*, and also on the harsh practical example provided by the United Kingdom's return to the gold standard in 1925 (Eichengreen 1992). He argued that, for many countries, sufficient competitiveness would not be assured if the world returned to a gold standard after the war. Such a standard would require that any country with balance of payments difficulties, of the kind which Britain was likely to have, would need to rely on downward adjustment of its wages and prices in order to make its goods sufficiently attractive in world markets. Keynes judged that, in the political climate of the post-war world, such wage and price adjustments might not be possible. Nevertheless, because of the exchange rate instability of the early 1920s and the 1930s, he also showed no enthusiasm for floating exchange rates. The need for something different was discussed in much detail over the next two years with Harry Dexter White and others from the United States (Keynes 1971–88, vol. 25, pp. 338 ff.), including during a visit that Keynes made to Washington in 1943.

The analytical content of these immensely difficult negotiations is explained in Meade, James Edward, and is discussed in more detail in Vines (2003), which draws on the wonderful historical account by Skidelsky (2000). Skidelsky makes clear that Keynes was propelled in these discussions by the knowledge that the generous provision by the United States of wartime funding to the United Kingdom ('Lend Lease') had put the United States in a position in which it would be able to dismember the British Empire after the war. Keynes, who had been accustomed to Britain managing the global economy, wanted to create a new global order in which prospects for Britain remained acceptable, even although global economic hegemony would pass to the United States. He feared that difficulties in the balance-

of-payments adjustment process might impose, on deficit countries like Britain, an obligation to deflate demand below full employment, something which might not be matched by symmetrical over-expansion by surplus countries, and might thereby create pressures towards global deflation. This is why he wanted his Clearing Union to be able to create global liquidity. (Like a bank, it would 'clear' the overdrafts which countries could obtain from it.) He differed in this view from Harry Dexter White, who feared an outcome in which liquidity would be so freely available that there would be a great post-war worldwide inflation.

What emerged at Bretton Woods was a global system of pegged but adjustable exchange rates, to be overseen by an International Monetary Fund. The currency system was to have three major features. First, each country would establish a *par value* for its currency in terms of gold or dollars. Second, all exchange controls would be removed for current-account transactions and all currencies would be freely convertible into dollars, although controls on international capital flows would remain in place. Third, dollars would be freely convertible into gold. Thus, the system was to be a 'gold exchange standard'; it would differ from a gold standard in being a club rather than a unilateral pegging arrangement, and in allowing for occasional exchange rate changes.

The IMF would do two things in this system. First, exchange-rate pegs would only be adjusted if the approval of the IMF's Executive Board had been obtained. That approval would not be given unless there were deemed to be a 'fundamental disequilibrium'. This term was imprecisely defined, but it meant a situation in which an exchange rate was not at a level that would ensure that exports could equal imports at full employment. This kind of test was designed, with the 1930s in mind, to prevent countries pursuing a 'beggar-thy-neighbour' devaluation of their currencies so as to steer towards full employment by 'stealing' jobs from other countries rather than by expanding expenditure at home. A country with longer-term difficulties would be declared to be in 'fundamental disequilibrium' and would be expected to devalue its currency by an appropriate amount after consulting with the Fund and getting the required approval. Similarly, a country with an excessively large and sustained balance of payments surplus would be expected to revalue its currency.

Second, the Fund would be set up like a credit union, into which members would place deposits; a country in temporary balance of payments difficulty rather than 'fundamental disequilibrium' would be able to draw on a short-term basis from the Fund to help it address the problem. It was thought that these loans would be repaid quite rapidly (that is, within three to five years), since more fundamental difficulties would be addressed by exchange rate adjustments. Each country in this credit union was to be given a 'quota', based on a nonlinear equation that took account of a country's national income, its international trade, and its official reserves; services, other external current account transactions, and a measure of volatility were further added to the quota formula in the 1960s. The quotas would define each country's capital contribution, its borrowing entitlement, and, in aggregate, the Fund's lending capacity. The US quota was initially about 20 per cent of the total (less than would have been implied by a strict calculation based on the variables noted above), and originally the United Kingdom had, by design, the second largest quota. This was *not* like Keynes's Clearing Union, and Keynes was dismayed at how little the Fund would be able to lend (see Vines 2003). There have been a number of substantial increases in total quotas under regular quinquennial reviews, but they have not grown in such a way as to keep pace with the expansion of the world economy and international financial flows. In addition, as the relative size and importance of countries have changed, there has been a need to adjust both quota shares and the factors used in the calculation of these quotas. Both of these types of adjustment have been politically difficult; a (small and interim) adjustment for four emerging-market countries (China, Korea, Mexico, and Turkey) happened in September 2006.

The quota system partly determined the relative voting entitlements of countries on the

Executive Board of the Fund. It seemed obvious, for a credit union to which money had been contributed, to make voting power depend partly on the amount contributed, and on the amount which could be borrowed at a time of difficulty, rather than using a one-member, one-vote system of governance like that adopted at the United Nations. However, there were also a number of 'basic votes' allotted equally to all members, whose effect was to mitigate a little the voting power of large countries.

The Fund's Articles and their subsequent amendments established that a member is allowed to borrow up to a certain proportion of its quota as of right, without policy conditions. This amount was referred to as the 'reserve tranche'; it was equal to 25 per cent of quota and corresponded to the amount that a member had paid into the Fund in hard foreign currencies. Beyond the reserve tranche, a country had an option to borrow up to four 'credit tranches', each of which represented 25 per cent of quota. Access to the first credit tranche was relatively easy; borrowing under the subsequent or 'upper' credit tranches was normally made available through what were (and still are) described rather quaintly as 'stand-by arrangements'.

Consequences

The international monetary system followed only imperfectly the intentions underpinning the Bretton Woods system, and only until 1971. (See de Vries 1976.) Current-account convertibility, for most European currencies, was not achieved until 1958 (the year after a large US current account deficit). There was a reluctance to alter exchange rates even in the presence of 'fundamental disequilibrium'. And the Fund was unable to stop France from implementing a multiple currency system in 1948. One major currency, the Canadian dollar, floated from 1950 to 1962 and the Fund acquiesced in this. The Fund ratified British devaluations in 1949 and 1967 at short notice (though it was closely involved in discussions in the second case). It had little influence on US policies – and has had little influence ever since. It played virtually no role in the later US decision to end gold convertibility in August 1971, a decision which brought the Bretton Woods system crashing down. And it had limited influence on the policies of the principal surplus countries in the 1960s. On the other hand, the Fund did have a role in the exchange rate realignments of other currencies that took place in 1949, 1967 and 1971 as a result of the sterling and dollar devaluations, seeking to ensure 'orderly adjustment'. The most important point is that the IMF had an influence mainly through the conditions it could impose on those countries (such as the United Kingdom in 1976) which needed its funds.

When the Fund began providing stand-by arrangements in 1952 they were typically of short duration and did not feature any conditions. This may seem surprising now, given the close association in the popular imagination between conditional lending and the IMF. Policy conditions were first added to Fund- supported programmes in 1954, partially in light of the increase in the size of borrowing under stand-by arrangements, as compared with first-credit tranche financing. Quantitative targets or 'performance criteria' followed in 1957, in order to provide a clear baseline for policymaking under IMF-supported programmes, and an objective yardstick by which the effects of these policies – and the possible need for further adjustments – might be assessed. They were calibrated using the Fund's financial programming framework, developed by Polak (1957), and came to be a nearly universal feature of Fund-supported programmes by the mid-1960s. (See IMF 1987; 2004a; Mussa and Savastano 1999.) This combination of policy or 'structural' commitments and quantitative performance criteria came to characterize the 'conditionality' attached to IMF lending from the 1960s to the present. This was justified – then as now – not so much as a way of collateralizing IMF lending, and guaranteeing a turnover of the IMF's funds, but rather as a means of ensuring the viability of Fund-supported programmes and the quick adjustment of countries in crisis back to a balanced growth path.

The period from 1945 to 1971 was one of extraordinary dynamism (a 'golden age'): it was a time in which Europe and Japan were first rebuilt after the war and then proceeded to

catch up with the United States. The Bretton Woods system appears to have played a part in ensuring that this happened. In this system, the Fund was helped by the World Bank, whose role was to lend money for longer periods than the Fund, first for reconstruction after the war, and then, later on, to help finance development. (Keynes once helpfully remarked that in order to comprehend the Bretton Woods institutions one has to understand that the Fund is a bank, and the Bank is a fund.) The purpose of this World Bank lending was to enable these countries to borrow abroad (in a world in which there was little international mobility of private capital), to run balance of trade deficits, to invest, and to grow – with the expectation that the borrowing would then be repaid out of the increased export proceeds that investment and growth made possible. In addition, a conference in Geneva in 1947 established the General Agreement on Tariffs and Trade (or GATT) to supplement the Bretton Woods system by encouraging the growth of international trade. The GATT's role in promoting the liberalization of trade restrictions supplemented the Fund's role in promoting the liberalization of exchange restrictions on current account transactions. In due course, a series of GATT 'rounds' brought about tariff reductions, which helped to create markets for exports as countries expanded. With high employment, with balance-of-payments deficits dealt with as described above, and with many countries growing by exporting, there were clear incentives for most countries to support trade liberalization. That, in turn, made exports and imports more sensitive to exchange-rate levels and so made balance of payments adjustment easier to achieve by exchange-rate adjustments. Yet, these linkages between different aspects of the overall post-war policy framework are difficult to pin down empirically. This explains why economic historians still differ in their view as to how important the Bretton Woods system actually was in sustaining the golden age of growth observed in the 1950s and 1960s. (See Matthews et al. 1982; Matthews and Bowen 1988; Temin 2002; papers in Eichengreen 1995; and Eichengreen 2007.)

Breakdown and Reconfiguration

Up to the 1960s the growth of gold reserves had been slow, and the need for additional international liquidity was increasingly met by the use of the US dollar as a 'reserve currency'. This led to calls for the IMF to create a more multilateral way to augment official reserves. The IMF's Articles of Agreement were eventually amended in 1969 to allow the Fund to create 'special drawing rights' (SDRs) that would act as the Fund's unit of account and which could be used as a source of credit for member countries. (See Corden 1983a; Boughton 2001.)

In the 1960s, imbalances also began to emerge: by the latter part of the decade, the United States had a large balance of payments deficit. A belief emerged that the dollar price of gold might rise as economic growth in Europe and Japan weakened the US dollar's role as anchor of the Bretton Woods system. In 1968, central banks ceased their efforts to control the dollar price of gold in private markets, which meant that the prevailing fixed price of gold applied only to central bank dealings. The market price of gold rose: in August 1971, following a massive speculative attack on the dollar, the United States ended the gold convertibility of dollars held by central banks and, as a result, the entire gold exchange standard broke down. A reluctant movement from a pegged exchange-rate system to a system with floating exchange rates followed. This outcome can best be explained by three sets of factors. (See Corden 1993.)

First, many countries were unwilling to adjust the exchange rates for their currencies in the face of fundamental disequilibria. It was particularly problematic that the core country, the United States, behaved in this way. Because US productivity growth lagged behind that of the countries which were catching up with it, the trade position of the United States was at risk by the late 1960s. In addition, the United States fought the Vietnam War and launched its 'Great Society' programmes at the same time, without adequately raising taxes. The result was a large balance of payments deficit for the United States, the correction of which required both real exchange rate depreciation

and restraint of domestic expenditure. Neither of these actions was forthcoming.

Second, the growth of international capital flows – which was in part a result of the international stability associated with the golden age – helped to undermine the system. As first demonstrated by the 1967 sterling crisis, it was no longer possible for the IMF and national governments to set exchange rates without reference to the forward-looking perceptions of private markets about what sustainable exchange rates might be. With increasingly mobile capital, once a suspicion was generated that that there would be (or might need to be) a devaluation of a country's currency to preserve external balance, speculation could make it difficult or impossible for central banks to defend an existing rate. By 1971, the balance of payments deficit of the United States had caused a large build-up of mobile dollar holdings in offshore or 'Euro-dollar' accounts. These funds were used to finance the speculative attack on the dollar in 1971.

Third, the Keynesian macroeconomic policy framework established after the Second World War contained no clear responsibility for preventing inflation. Although there were periods of (generally unsuccessful) price controls or 'incomes policy', the seeds of incipient inflation were sown by this omission. Eventually, tensions generated by the oil price shock of 1973, and by the period of undisciplined inflation which followed it, led to more than the collapse of the Bretton Woods system. The entire structure of Keynesian, interventionist, high-employment policies, which had been at the centre of the post-war policy architecture, came tumbling down, both in the United States and in Europe. For the ten years after 1971, macroeconomic policy was in a state of worldwide disarray.

The great inflation of the 1970s led to significant movements in the real exchange rates between countries, which killed nearly all of the (many) attempts made at the time to reconstruct an international monetary system with pegged exchange rates. (See Williamson 1977.) There was only one lasting, partial, attempt to reconfigure such a system, in Europe, which led to the European Monetary Union.

For a period of time it appeared that the Keynesian approach to macroeconomic policy might be replaced by monetarist policies of a non-interventionist kind. But this alternative proved unsuccessful. Instead, with great difficulty, activist macroeconomic policies were reconstructed by the 1990s within inflation-targeting regimes, in which an inflation target was pursued through interest rate changes. This new system quickly came to be allied with a system of floating exchange rates in which there was a high degree of international capital mobility. In this new set-up, a floating exchange rate would help to stabilize demand, and movements in the exchange rate would become an important part of the process of inflation control. If a country suffered from a shock which raised prices, then its monetary policymakers would set higher interest rates, and the nominal exchange rate of the country would appreciate. This would reduce net exports and import costs, and so inflation.

As a result of this reconfiguration of policy assignments, a second revision of the Fund's Articles of Agreement was made in 1976 and came into effect in 1978. At Bretton Woods, the Fund had been set up to manage a pegged exchange rate system. But it came to be realized that a country cannot have, at the same time, an independent monetary policy, capital markets which are open to the rest of the world, and a pegged exchange rate. (These three things, taken together, have become known as an 'impossible trinity'. The reason that these things cannot occur together is to be found in the Mundell–Fleming macroeconomic model, which was developed by Fleming and Mundell, at the IMF, in the early 1960s.) As a result, the Fund's revised Articles ratified a new form of international monetary system in which a country did not have to establish a par value for its exchange rate, but could instead have exchange rate arrangements of its own choice.

Since 1978, the Fund has gradually been drawn into new roles, in support of this revised, and more flexible, system. As described in the introduction, its work now has three aspects. First, the Fund's Articles, as revised in 1976, require it to exercise surveillance and influence over macroeconomic policies, and to monitor and guard against the

development of unsustainable conditions that could lead to financial crisis. The Fund still lends to countries in balance of payments difficulty, and its second activity has been to do this for emerging-market economies and for 'transition economies' moving from central planning to market-based systems. More than this, the Fund helps such countries to deal with, and to prevent, the financial crises that have afflicted a number of them. Third, the Fund has lent money to the poorest developing countries, which generally do not have capital-market access. In these cases, Fund lending has often been indistinguishable from other long-term concessional development assistance, and the Fund's main distinctive contribution has been to work with central banks and finance ministries in crafting credible macroeconomic frameworks that can elicit further support from aid donors. We consider each of these three activities in turn.

The IMF and Policy Surveillance

Countries that are creditworthy, and which have access to highly mobile international capital under floating exchange rate regimes, no longer need to borrow from the Fund in the way they did when the Fund was first established. Such countries can adjust to balance of payments disequilibria through exchange rate movements, supported by foreign borrowing from sources other than the Fund. (See Corden 1983b, and Dam 1982). At the time of writing, no advanced country had agreed a borrowing arrangement with the Fund since the substantial stand-by arrangements with the United Kingdom and with Italy in 1976. Fund lending is only required at a time when a country ceases to be perceived as clearly creditworthy, something which, as of mid-2007, had not happened in industrial countries since 1976. This was true even at the time of the crisis of the Exchange Rate Mechanism of the European Monetary System in 1992. The Fund did not at that time provide financing to assist Sweden, Italy, the United Kingdom, or France in a defence of their currencies. When crisis struck, these countries (eventually) allowed their currencies to float downwards, rather than using lending from the IMF to defend further their exchange rates.

Nevertheless, a world with a high degree of international capital mobility is not without difficulties. In such a system, the spending decisions of nations can move away from permanently sustainable positions for very long periods of time, an outcome with an external current account deficit (or surplus) offset by an external capital account surplus (or deficit). The 'global imbalances' that can result have, as of mid-2007, been substantial at three points of time since the 1960s. In the late 1960s, as we have seen, the US ran a large current account deficit; current account surpluses of a number of European economies and of Japan, which, as noted above, were engaged in a process of export-led growth and 'catch-up', were the 'other side of the coin'. Nearly 20 years later, in the early to mid-1980s, President Reagan increased defence expenditures and cut taxes. Tight monetary policy was used to restrain demand in the United States, which caused the dollar to appreciate, and the result was a large current account deficit. Japanese current account surpluses were on the other side of this coin. Twenty years later, in 2007, the United States was again running a large fiscal deficit and an (unprecedentedly) large current account deficit; and again Japan was running the corresponding current account surpluses, along with China, other emerging-market economies in East Asia and elsewhere, and a number of oil-producing countries.

These global imbalances reflect decisions by countries to de-link income and spending over time. Of course, such 'intertemporal trade' can be welfare-improving. But such imbalances might instead reflect an urge by a deficit country to spend beyond its means. This was clearly the case for the United States in the late 1960s and the mid-1980s, and might also be the case from 2000 (and especially from 2005). Conversely, these imbalances might also partly reflect a desire by some countries to maintain their currencies at artificially devalued levels against the US dollar, in order to grow quickly through a process of export-led catch-up. This is something which, at one time, would have been called 'beggar-thy-

neighbour' behaviour of the kind which the IMF was established to prevent. As noted above, one can argue that this may have been what was done by western Europe and Japan in the late 1960s. Some commentators have argued that a number of emerging- market economies in East Asia, and elsewhere, were behaving the same way in the early 21st century (Dooley et al. 2003; Roubini and Setser 2005). These commentators, in recognition of the parallel, suggested that we were living under a 'Bretton Woods II' regime.

But global imbalances eventually unwind. They must do so if countries are eventually to repay what they owe. In 1971, global imbalances led to crisis, and to the collapse of the Bretton Woods financial system. By contrast, the imbalances of the mid-1980s were resolved in an orderly way. (See Eichengreen 2004; Eichengreen and Park 2006; Corden 2007; Joshi et al. 2006; Williamson 2006.) Such orderly adjustment requires the deficit country to cut expenditure, and its currency to depreciate significantly (unless it grows its way out of difficulty). It also requires, in addition, that expenditure in surplus countries expands so that global expenditure is maintained, or, if this does not happen, that global interest rates fall so that global expenditure is stimulated by other means. If all of this happens, as it did in the late 1980s, then the benefits of intertemporal separation between spending and income may not be diminished by the costs of an adjustment crisis.

There are four main ways in which the existence of the Fund helps global imbalances to unwind in an orderly manner.

First, ever since the second amendment of the Fund's Articles described above, the Fund has been required to exercise 'firm surveillance' over the exchange rate and macroeconomic policies of its members. As a result, the Fund regularly sends to each country an 'Article IV mission' whose purpose is to review the country's macroeconomic policies. This is done annually for most countries, and at interludes of up to 24 months in countries with active Fund-supported programmes. (For such countries the Article IV cycle is elongated since policies are reviewed frequently in the context of semi-annual or quarterly programme reviews.) All aspects of macroeconomic policy are considered on these occasions. Following the emerging- markets crises of the 1990s and early 2000s, the Article IV consultation process has been supplemented by detailed review of countries' financial sectors under the World Bank and IMF's joint Financial Sector Assessment Program (FSAP).

Second, the Fund provides a vast amount of published information and analysis, both about the world economy and financial system in general and about particular countries. The Fund's biannual *World Economic Outlook* provides a forecast for the world economy, and analyses multilateral and regional issues; this report is supplemented by *Regional Economic Outlooks*. These products are based in part on Article IV consultations and would not be possible without that process. The Fund also publishes a biannual *Global Financial Stability Report* which monitors markets, and several statistical publications that compile economic and financial data supplied by member countries, including *International Financial Statistics*.

Third, the Fund plays an important role in keeping the governments of all members in touch with developments in other countries and globally. The Article IV missions to the largest economies (and the related research, published in *Selected Economic Issues* papers that are companions to the Fund's Article IV staff reports) are particularly important in helping to keep governments informed of policies and developments that are likely to affect the world economy as a whole. Additionally, the Annual Meetings of the Boards of Governors of the IMF and the World Bank enable an informed exchange of ideas between countries, as do the Spring Meetings. The Fund thus provides a valuable global information network.

Finally, the Fund has also created a valuable global human network. Fund staff are of high quality, something which is necessary since they have to deal with senior officials in many countries. The offices of Executive Directors of the Fund in Washington act as valuable means of communication between the member nations of the Fund. And in many national capitals a large

number of public servants and elected officials have served on the Fund staff earlier in their careers, or have been located in Washington as Executive Directors at the Fund or as members of staff in Executive Directors' offices. This experience has made many decision-makers more internationally minded than they might otherwise have been.

Nevertheless, some have argued that the Fund's 'firm surveillance' is not firm enough. Arriazu et al. (1999) discuss the impact of Fund surveillance, country by county, in the Article IV consultation process. They note that, although these consultations have been 'taken seriously', it does not appear that these reviews by the Fund have had more than an occasional impact on national policy decisions in some countries. A more recent assessment of Article IV consultations by Meyer et al. (2004) reaches similar conclusions. When an Article IV mission goes to a country that does not borrow from the Fund (and which therefore does not require the Fund's *imprimatur* in order to obtain loans from other official creditors or from banks), the mission is usually relegated to a mainly advisory role, for which 'surveillance' may be too grand a label. But this *de facto* situation is not inevitable, since the *de jure* position of the Fund is that it should assess and appraise as well as advise. Goldstein (2006) asserts that there are gaps in the current practice of bilateral surveillance and argues, in particular, that the Fund's dealings with China in the early 21st century have not been satisfactory in addressing and effecting remedies for exchange rate misalignments. He further observes that the Fund's Managing Director has only rarely used the power granted to him by the 1977 and 1979 Board decisions on ad hoc and 'supplemental' consultations with members to address cases where a country's exchange rate policies appear inconsistent with the exchange rate principles of the Fund's Articles. (See Boughton 2001.)

It is important to note that these critics do *not* seek policy changes from countries, in the interests of the greater good, that such countries would find unattractive if left to make policy choices on their own. That is, it is not suggested that the Fund could enforce a 'cooperative' outcome in macroeconomic policymaking when countries would prefer a different selfish, or 'Nash,' outcome. (This difference between Nash and cooperative outcomes was much discussed in the 1980s literature on policy coordination, summarized by McKibbin 1997). Instead, it is argued that the Fund could enable cooperative outcomes, so that any adjustments in countries' policies that need to happen in the face of global imbalances might happen in the right sequence rather than in a disorganized manner. The capacity to enforce even this modest form of coordination might occasionally be important in the adjustment processes. (See Kumar 2006; Wolf 2005, 2006; Joshi et al. 2006.)

There was action of this kind under the Plaza Accord of September 1985, although it was not coordinated by the Fund. At this time, the finance ministers of the world's five largest national economies agreed that the value of the dollar needed to go down. They also arrived at some (rather general) agreements on the monetary and fiscal policies that would be needed in order for this fall in the dollar to be achievable, and announced coordinated intervention in foreign-exchange markets to help bring it about.

To act effectively in this way requires the Fund to come to terms with the difficult tension between its strengths as a universalist institution and the need, on occasion, to bring together a more limited group of players. But it is an objective of the Fund's current Medium-Term Strategy that it should provide such a forum (IMF 2005b). The Fund's Multilateral Consultation on global imbalances began by consulting with the United States, the European Union, Japan, China and Saudi Arabia, and it reported on its findings in April 2007. This work ran in parallel with similar discussions at summit meetings of Heads of Government of the Group of Eight Countries (or G8), and at meetings of the finance ministers and central bank governors of these countries. The G8 consists of the United States, Russia, Japan, Germany, Britain, France, Italy, and Canada. This is a powerful collection of countries, but it is not clear that these G8 meetings have had the right participants to deal with the global imbalances of the early 2000s. China and India have not been members

I

of this group (though they have been observers), nor have many of the major oil-producing economies; by contrast, Canada and Italy, while committed to the G8 process, have been perhaps too small to contribute substantially to coordinated efforts to unwind global imbalances. The Fund may therefore have more to offer than such G8 gatherings, since the Fund can act as a locus of coordination amongst subsets of its membership, convening small groups of countries to deal with particular problems.

Nevertheless there are three reasons why further progress may be slow on this front.

First, in the words of the IMF's Independent Evaluation Office (IEO) (IMF 2006a, p. 2), 'As a result of its ... [country-by-country] orientation, multilateral surveillance has not sufficiently explored options to deal with policy spillovers in a global context'. Pursuing this theme, Mervyn King, Governor of the Bank of England, made it clear (King 2006b) that more effective multilateral surveillance would require: (i) that countries made clearer commitments about their objectives for macroeconomic policies (that is, fiscal, monetary and financial); (ii) that the Fund's Article IV and the *World Economic Outlook* processes focused more transparently on cases when these policy commitments, and the countries' policy actions, are not globally consistent; and (iii) that this process also transparently demonstrated the negative spillover effects that come from such lack of consistency and proposed actions to reduce such negative spillovers. But, given the limits to the precision of what we know about the international economy at any given time, doing this would be difficult. And it should be noted that the Fund's management issued a rejoinder to the 2006 IEO report which explained this difficulty.

Second, there may well be governance limitations on such firm surveillance. As of 2007, Article IV consultations were not finalized by the Fund Staff sent on the Article IV mission, but by the Fund's Executive Board, whose views were conveyed to the authorities of the country concerned after discussion at the Board. It is possible that this has compromised the space for missions to assess and appraise frankly. If the process of IMF surveillance were made more independent of the IMF's Executive Board, then this might allow clearer messages to be delivered to the Fund's member countries. As against this, the messages might then lose political weight because they would no longer be seen as the views of the global community represented in the Executive Board.

Third, and fundamentally, the Fund is not an agent of a sovereign state in the way that central banks (except the European Central Bank) are, however 'independent' these central banks may be. As a result, the Fund has no actual instruments of its own with which its recommendations on global cooperation can be implemented. It must always rely on being able to persuade its members to act.

The IMF and Crises in Emerging Markets Since 1980

In the mid-to-late 1970s, after the rise in the price of oil in 1973, funds flooded from oil producers on to the international capital market and flowed to middle-income countries. The early to mid-1990s saw a further massive surge of private capital flows into emerging market economies, and this was repeated in the mid-2000s. The economic benefits of such international mobility are obvious: if capital flows from relatively rich to relatively poor countries, and if the rate of return is high in poor countries, the potential gains are high for both borrower and lender. But such funds are not always used well, the volatility of these flows can be very high, and they can create dangerous mismatches in the maturities and currencies of assets and liabilities. Indeed, these flows contributed to three major waves of financial crises, in Latin America, East Asia and Russia, something which called into question the stability of the entire international financial system. Across these regions of the world, the IMF has been required to help prevent such crises through surveillance. It has also been required to assist in the orderly workout of crises, through lending and through ongoing engagement in the development of macroeconomic policies in the countries which it assists. We explain how the Fund's activities

have evolved in these emerging-market economies, and how its role has broadened. We do this by examining the three generations of emerging-markets crises that occurred from the early 1980s onward.

The Latin American Debt Crisis: A 'First-Generation' Crisis

Oil money, facilitated by loans from international banks, financed a spending boom in Latin America and elsewhere during the 1970s. This led to a rapid increase in foreign debts (Little et al. 1993) in countries which were not in a position subsequently to adjust and service these debts. In due course, significant balance of payments problems emerged when, in 1980–82, real interest rates rose, driven by tight monetary policy in the United States and by a world recession which worsened the terms of trade for many emerging-market economies. These countries rediscovered the truth of what Keynes had maintained 40 years earlier: adjustment to external difficulties requires both good budgetary control and an appropriately competitive real exchange rate (Corden 1990; Little 1993). This turned out to be something which many policymakers in Latin America, and elsewhere, were unable to engineer, and monetized fiscal deficits led to reserve losses, uncontrolled devaluations of currencies and inflation, and difficulties in meeting foreign-currency-denominated debt obligations. Currency and debt crises were triggered more or less mechanically as macroeconomic fundamentals drove reserves down to critical levels, resulting in what has become known as a 'first-generation' crisis.

Although Latin America is most closely associated with the debt crisis of the early 1980s, other countries, including Morocco, were also involved. The crisis placed the IMF at the centre of the world stage in a way which made it more prominent than it had ever been under the Bretton Woods system. The Fund played four roles. First, it offered financial support with stand-by arrangements and other lending facilities. Second, the Fund came to define the broad envelope of resources that a country could be expected to devote to meeting its residual obligations under a debt rescheduling. In turn, the Fund, together with the United States and other bilateral creditors in the Paris Club, pressed creditor banks to reschedule debts and to engage in 'concerted lending' programmes, threatening to provide no support for indebted countries if banks did not cooperate, and, hence, making defaults more likely. Third, the Fund's advice and conditionality, together with that of the World Bank, had significant effects on indebted governments' policies: they were encouraged to undertake growth-oriented structural reforms to escape from their debt problems. Fourth, the Fund's reports and conditionality provided the 'seal of good housekeeping' on the basis of which banks and bilateral creditors could justify rescheduling existing debt and providing new funds.

This use of the Fund, and the broader strategy surrounding it, is usually associated with James Baker, then Secretary of the US Treasury. It was a success only to the extent that it made the financial crisis manageable. The strategy avoided explicit debt reduction and insisted that indebted countries meet their obligations, although over an extended period of time. (This lengthening of the repayment profile did, of course, lead to some reduction in the net present value of debt.) Such an approach was advocated by the governments of major industrialized countries, especially the United States, that were concerned about systemic risks to their own banking systems arising from widespread write-downs of debt. The Fund was criticized in some quarters for agreeing to this strategy and for acting as an 'enforcer' of debt service on behalf of private banks.

A policy shift took place in 1989. Under the Brady Plan, also initiated by the US administration, the Fund and the World Bank provided encouragement and some financial support for debt reduction programmes for those countries (notably Mexico) where major policy reforms were being undertaken. The shift from the Baker Plan to the Brady Plan represented a tilt in favour of debtor countries relative to creditor banks. But this came only after a long period in which these banks were able to rebuild their balance sheets, thereby putting them in a position to weather debt restructuring. The US Treasury induced creditors to grant write-downs to debtor countries by

collateralizing the debt that emerged from these restructurings. The Fund backed up this carrot by concluding financing packages with debtor countries before the terms of debt reschedulings had been determined: a practice that came to be known as 'lending into arrears'. This acted as a stick to weaken creditor leverage in the negotiation process, and it also greatly strengthened the role of the Fund in debt work-outs since, during the negotiations, Fund staff came to play a major role in influencing debtor countries' macroeconomic policies.

The Mexican 'Tequila' Crisis: A 'Second-Generation' Crisis

The Latin American debt crisis of the early 1980s had been caused by *public*-sector overspending. But in 1994 something new happened. A major financial crisis, caused by the outflow of *private* capital, of the kind which had brought down the Bretton Woods system in 1971 and the European Monetary System in 1992, happened in Mexico. The Mexican crisis was different from the Latin American turmoil of the 1980s in that it was set off not just by fundamental weaknesses, such as unsustainable fiscal and current account deficits, but also by currency mismatches on the public-sector balance sheet. (See Calvo and Mendoza 1996.) These caused a 'second-generation crisis' in the form of a self-fulfilling currency run. This crisis presented new challenges for the IMF since it marked the first of a series of crises in emerging markets that originated in the capital account, rather than the current account, of the external balance of payments. The IMF was called on to assist Mexico despite the fact that its Articles of Agreement provide it with only limited jurisdiction over capital account issues.

Mexico had implemented a comprehensive reform programme in the early 1990s, which included financial liberalization and the completion of the North American Free Trade Agreement (NAFTA) in 1993. This led to a surge in investment financed mainly by foreign capital flows. The result was a large (real) overvaluation of the peso and a very large current account deficit. Initially, the government maintained prudent fiscal policy. But during 1994 many began to question the sustainability of the exchange rate, the fiscal position and current account deficit. By December 1994 there was a massive reversal of capital flows, and the peso plummeted. The consequences for Mexico were severe: inflation rose from 7 per cent in 1994 to 35 per cent in 1995; and GDP fell by 6.2 per cent in 1995 compared with a growth rate of 4.4 per cent in the preceding year.

The pain inflicted on Mexico by private investors led to a view that pegged exchange-rate regimes are unviable everywhere, not just in advanced industrial countries. (Mexico had a 'crawling peg' at the time.) And in Mexico there was a new emerging-market feature. Much of the Mexican government's debt was denominated in US dollars (for example, the *'tesobonos')* because of the difficulty and high costs of borrowing in local currency; much of the government's revenue stream, by contrast, was peso-denominated (although oil revenue was denominated in dollars). This mismatch meant that the collapse of the peso led the government to the verge of default in early 1995.

The Fund played a critical role in stabilizing the crisis. In particular, drawing on financing from bilateral creditors, it coordinated assistance, mainly from the United States, that totalled more than five times Mexico's quota entitlements at the IMF. After a significant real devaluation of the peso and fiscal correction, exports rebounded, the economy grew, although only slowly, and Mexico earned enough foreign exchange to repay the exceptional financing that had been provided to it during the crisis.

Some subsequent analyses (see, for example, Calvo and Goldstein 1996) were critical of the IMF's role in both surveillance and in crisis management for Mexico. But the arguments cut both ways.

On surveillance, it was claimed that IMF reports prior to the crisis placed insufficient emphasis on the vulnerabilities of public-sector and financial-sector balance sheets to the possibility of a run on the currency. Some authors argued that the Fund should have been more frank in conveying its views on macroeconomic and exchange-rate policy to its members, and that it should publish these appraisals. But there may

well have been inadequate provision of information by Mexico to the Fund, as well as to the public. In particular, it appears that incomplete data may have been provided on official international reserves and liabilities (although the Mexican authorities disagreed with this claim when it was made). As a result, following the Mexican crisis, the Fund began a drive to get countries to sign on to transparency standards, such as the Fund's Special Data Dissemination Standards (which were established in 1996; see Fischer 2004, p. 127). Additionally, the Fund began the practice of publishing Board documents, except when the authorities of a country objected. But this heightened focus on transparency left the Fund unclear on whether it should assist countries confidentially to prevent crises or spur corrective action by bringing bad news to the market. Given the sometimes self-fulfilling mechanics of second-generation currency crises, solving this dilemma is critical in defining the future role of the Fund in crisis prevention.

On crisis management, no clear conclusions emerged, either. *Ex post* it appeared that the private sector should have been prepared to lend short term to the Mexican government in the way that the IMF and the United States did. Overcoming such a market failure is surely a role of the IMF and national governments, and giving the IMF the capacity to provide such big loans seemed important to many observers. From this experience, Sachs (1995) concluded that the Fund should be given an explicit international lender-of-last-resort capacity, well beyond that formally possible under its 'credit-union' status, so as to enable it to be ready to respond forcefully and quickly to emerging crises, as it had done in the Mexican crisis. (See also Fischer 1999.) With such firm IMF action, currency crises could be contained as liquidity crises rather than becoming solvency crises. Indeed, it appears that the combination of large-scale IMF financing, combined with significant adjustment by the authorities, prevented the development of a solvency crisis in Mexico. However, some authors began to warn that, if the IMF always acted as a lender of last resort in the face of crisis, then this might create moral hazard on the part of lenders to emerging markets, who might expect to be able to lend virtually risk-free with any possibility of default prevented by IMF action. (The Fund-led bailout of *tesobonos* holders strengthened these fears.) These critics suggested that efforts be made to make sovereign debt rescheduling easier and more orderly (Eichengreen and Portes 1995), thereby containing the threat of creditor moral hazard.

The Asian Financial Crisis of 1997–98: The 'Third Generation' of Crises

Two and a half years later these issues re-emerged in Asia, in a crisis which interrupted a long period of sustained economic growth financed by exports and foreign capital inflows. Unlike the earlier Latin American debt crisis, or even in Mexico, fiscal profligacy played no *explicit* part in the East Asian crisis. But there were two other main policy failings. (See Bluestein 2001; Corbett and Vines 1999a, b; Corbett et al. 1999.)

First, much more than in Mexico, an underdeveloped financial system and over-protected financial sector in some Asian economies meant that the private sector had to rely on borrowing, rather than equity issuance, to raise investment funds. As a result, firms became highly leveraged, but banks continued to lend because they were underpinned by *implicit* government guarantees. When growth slowed, as it first did in Thailand in 1996, and then in other East Asian economies, these banks were exposed to the inability of borrowers to repay loans.

Second, a further difficulty arose, as so many times before, from the existence of fixed exchange-rate systems in some East Asian economies, but with a new twist. Banks financed much of their domestic corporate lending by borrowing in foreign exchange from abroad, often at shorter maturities than those employed when they lent onwards in domestic currency. Very little of this borrowing was hedged as a result of the implicit guarantee on the exchange rate. As noted in the previous paragraph, the financial sector was already in difficulty after the initial slow down in growth in 1996. Currencies fell in mid- to late 1997 because of foreign investors' concerns about these difficulties; as a consequence,

widespread bankruptcies and potential bank failures loomed because of the unhedged foreign-currency obligations. Fear grew that fiscal systems would be unable to bear the cost of large-scale bank rescues (Irwin and Vines 2003).

The East Asian debacle marked the advent of 'third-generation' crises in which currency crises and banking crises are intimately intertwined – situations in which vulnerabilities in the private balance sheet can quickly translate into a public debt crisis.

As in Mexico, the Fund played a large part in resolving the crises. The IMF moved quickly to lend very large sums to Thailand, Korea and Indonesia. Nevertheless, there has been widespread criticism of the Fund's behaviour before and after the crisis. (See, for example, Stiglitz 2002.)

Two difficulties must be acknowledged in the Fund's *crisis prevention* work in East Asia. First, the Fund may have underestimated the risks associated with capital account liberalization. Second, the Fund may not have been firm enough in warning of the difficulties inherent in maintaining a fixed exchange-rate peg. Nevertheless, Thailand, for instance, was warned privately by the Fund several times in the year leading up to the 1997 currency crisis. The Fund, like some private-sector analysts, saw problems looming in Thailand, but its advice was not heeded.

Concerning the Fund's work on *crisis management*, there are three points to consider.

First, as the Fund has acknowledged in both its own reviews of the East Asian crisis and in the evaluations performed by its Independent Evaluation Office (IEO) (IMF 2003), its programmes may have placed too much emphasis on tightening budgets in countries that were already running prudent fiscal policies. Stanley Fischer, then the Fund's First Deputy Managing Director (FDMD), argues, however, that this approach was driven by a need to boost government savings to support the current account and provision for the impending cost of bank restructurings. (See Fischer 2004.) Furthermore, the credibility of an adjustment programme at a time of crisis may hinge on policy erring towards being too tight, in order to send a clear signal to markets. Once the scale of the economic downturn became apparent in East Asia and current account balances improved, Fischer argues that the Fund programmes shifted to addressing structural problems. (See also Corden 1999; Boorman et al. 2000.)

Second, monetary policy was also tightened in an attempt to defend currencies. There is an inevitable trade-off between raising interest rates in order to moderate exchange rate depreciations and lowering interest rates so as to ease the stress on both the banking system and on corporations that depended on domestic credit. Stiglitz (2002) argues that the tightening was too forceful. However, it does appear that this tightening was essential in order to stem capital flight. Nevertheless, this tightening was not followed by a concerted move to an inflation-targeting regime of a kind that might have allayed concerns of further depreciation. Hence, pressure on the region's currencies continued. And rather than stimulating recovery, these depreciations proved contractionary, at least initially, owing to their effects on external debt burdens. (See Krugman 1999.)

Third, the Fund did not have a mandate to declare 'standstills' on external debt payments during the crisis. In corporate bankruptcies, standstills force creditors to share in the burden of crisis and agree to reasonable debt reschedulings. In the context of a currency crisis, a standstill mechanism would similarly 'bail in' foreign private-sector creditors and then make reschedulings possible to reduce debt to sustainable levels. The fact that a standstill was not imposed in Thailand, Korea or Indonesia enabled creditors to race to get their assets out of these countries. Negotiations with foreign creditors to Korea and Indonesia did ensure some rollover of existing short-term lending, with effects similar to those that might have resulted from standstills. In both cases, however, negotiations were pursued too late and without sufficient coordination to maximize their impact (though they did stave off collapse in Korea). The only comprehensive brake on external payments was that imposed in Malaysia through the implementation of capital controls rather than a standstill by the government of Prime Minister Mahathir bin Mohamad in late 1998, a move that contravened the Fund's advice. But this was done

only after substantial capital outflows from Malaysia had already taken place.

Because the Fund lacked a mandate to impose standstills, it lent countries money in an attempt to allay the concerns of foreign creditors and to stem capital flight. Given the scale of the external capital-account movements in these countries, the size of IMF financing packages soared, especially after it became clear that smaller lending programmes would be unlikely to produce adequate results. In the case of Korea, the authorities of the IMF's large shareholder governments, notably the United States and Japan, also made a key decision to pursue a debt rollover plan and to exert moral suasion on creditor banks. These banks presumably realized that the alternative would have been partial default. The IMF played a useful role in facilitating communication among the different actors, in providing information, and in certifying that the policies to be pursued by the Korean authorities were appropriate. The IMF's Independent Evaluation Office writes, 'No single national government, nor any private sector institution, could have played this role as effectively' (IMF 2003, p. 115).

Although the Fund's work in Korea showed that the IMF could effectively manage a debt workout, its conduct elsewhere in the East Asia crisis had the effect of shifting the balance of power in debt workouts back toward creditors. IMF programmes did *not* reduce the debt overhang in Indonesia and Thailand. Instead, governments rescued banks and corporations by shifting their debt to the public balance sheet. Taxpayers in these countries still bear the burden of this debt. Rather than 'bailing in' private creditors, the Fund's handling of the crisis in these countries may have provided creditors with an even bigger bailout than they might have expected under the terms established in the 1990s' Brady Plan.

Partially out of dissatisfaction with this result, Anne Krueger, who followed Fischer as the Fund's FDMD in 2001, proposed a bankruptcy or standstill procedure for countries, the 'Sovereign Debt Restructuring Mechanism' (SDRM) (Krueger 2002). The US Treasury and financial markets both opposed this proposal out of a concern it would create unrestrained debtor moral hazard. Under what came to be known as the 'Taylor Doctrine' (after John Taylor, then US Treasury Under Secretary for International Affairs), the US government argued that countries should be left on their own to negotiate with their creditors. But this is only feasible when the number of external creditors is small, which for most countries has not been the case since the 1980s when external borrowing was provided mainly under loans from banks. To help remedy this problem, the US supported the introduction of 'collective action clauses' (CACs) in bond contracts with commercial creditors. These clauses prevent rogue creditors from holding out in restructuring negotiations in order to extract a premium from the bond issuer; they work by enforcing a restructuring if a pre-specified minimum proportion of creditors have agreed to its terms. CACs do not, however, provide a framework to guide the allocation of losses between borrowers and lenders, which is necessary in any restructuring. In the absence of a clear means of sharing these losses, it may prove impossible to renegotiate debt owed to commercial creditors. When faced with debt-servicing problems, debtor countries may then decide to borrow from official sources (including the IMF, whose debt is senior to other external liabilities and not reschedulable) in order to repay private sector creditors, as happened in Korea, Thailand and Indonesia. Since private-sector creditors are likely to believe that this will happen, the Taylor doctrine's approach, even when coupled with CACs, might promote creditor moral hazard, something which has been feared ever since the Mexican crisis. Thus, although the Taylor doctrine's approach has the virtue of minimizing debtor moral hazard, it appears to go in the opposite direction by promoting creditor moral hazard.

Default: The Russian and Argentine Crises

Russia. The fall of the Berlin wall in 1989 and the dissolution of the Soviet Union in 1991 enabled the IMF at last to become a (nearly) universal institution. In three years, membership increased from 152 countries to 172, the most rapid increase since the influx of African members in the 1960s. The IMF supported programmes in most former

Eastern Bloc countries and newly independent ex-Soviet Republics to help ease the transition to a market economy. The contribution the IMF made to the speed and relative smoothness of this transition is, perhaps, one of its most singular and least-heralded achievements.

Russia, however, got off to an inauspicious start under the first stand-by arrangement with the Fund in 1992. The IMF encountered intense difficulties in influencing the Russian leadership (Odling-Smee 2004). GDP fell for several years under the IMF-supported combination of macroeconomic stabilization and industrial restructuring. Although the IMF can claim credit for helping to instil some monetary discipline by the mid-1990s, the process took time, foreign direct investment remained low, tax collection was poor, and the fiscal deficit remained large. Growth in real GDP did re-emerge by 1997. But, following the onset of the East Asian crisis, the ruble came under speculative attack in November 1997. Pressure on the ruble was compounded by foreign investors' attempts to hedge their ruble holdings, as well as by a drop in the price of oil, which accounted for about one-third of Russia's foreign-exchange inflows.

Russia sought additional IMF financing in early 1998, but agreement on the terms of a new programme could not be reached owing, in part, to a failure by the Russian authorities to secure an increase in fiscal revenue. As a result, foreign investors began to unload Russian assets and about US$4 billion fled the country in the summer of 1998. By the time additional IMF financing was agreed in July 1998, fears of a devaluation led to such a pronounced sell-off of Russian securities that the authorities were forced to devalue the ruble and halt payments on both domestic and foreign debt.

Although the Fund is routinely criticized for providing cover for private capital flight from Russia in the first half of 1998, private investors who maintained faith that the Fund would rescue Russia sustained even greater losses when the ruble was devalued. This was perhaps the largest case to that point where the Fund stepped away from a floundering member, declared a solvency crisis, and let private creditors sustain substantial losses. It marked a different approach to the challenge of balancing creditor and debtor interests from that which the Fund had adopted in East Asia. And in some ways it set a precedent for the Fund's handling of the Argentine crisis in 2001.

Argentina. After a sustained period of hyperinflation in the 1980s, Argentina decided in 1991 to peg its currency, the peso, to the US dollar under a quasi currency-board regime at a one-to-one parity. Although the Fund cautioned that Argentina had neither the fiscal discipline nor the robust export sector needed to sustain such a system, it went along with the authorities' plans and supported their macroeconomic programme under a series of lending arrangements. By the late 1990s, Argentina was widely hailed as a model of successful economic reform as the rate of inflation fell to single digits and growth increased. In addition, the economy had successfully weathered the global turbulence caused by the East Asian crisis of 1997–8, and the Russian crisis of 1998.

But the seeds of the problems identified by the Fund back in the early 1990s were beginning to bear fruit by the end of the decade. Fiscal policy remained insufficiently tight owing to the lack of effective central government control on provincial borrowing, and this stimulated domestic demand for imports. Argentina's export sector remained too small to finance these imports, and its real exchange rate made its goods uncompetitive on regional and international markets. As a result, Argentina chose to borrow substantial amounts in US dollars to finance its imports. Brazil's decision to float the real in 1999 in response to pressure from the Russian crisis made it even harder for Argentina to compete under its quasi currency-board regime. The Argentine authorities allowed the peso to float in January 2002, and it quickly collapsed from parity with the US dollar to an exchange rate of nearly 3.9 to the dollar in June 2002. Output fell sharply, inflation reignited, the government defaulted on its debt, and the banking system was largely paralysed.

The Argentine debacle rightly cast several doubts on the Fund's conduct of both crisis prevention and crisis management in emerging markets. At the outset of the 1990s, the Fund proved incapable of resisting Argentina's arguably

doomed effort to impose its quasi currency board. Subsequently, the Fund endorsed Argentina's exchange rate peg in a series of programmes through the 1990s that coincided with an accumulation of macroeconomic vulnerabilities. When the regime became unsustainable in 2001 (or earlier), the Fund maintained lending until the end of that year in an attempt to save the peg. After the crisis, the Fund resumed lending to an insolvent Argentina in 2003 at the behest of the Executive Board, even although misgivings were expressed by the Fund staff. IMF lending ceased again later in 2003 and Argentina pursued an aggressive 'take it or leave it' strategy with private creditors. The Argentinean authorities achieved a roughly 75 per cent write-down on the country's defaulted foreign bonds, while leaving nearly US$20 billion in unexchanged bonds in default (IMF 2005a).

The Fund's experience with Argentina demonstrates at least four things. First, it can be very difficult for Fund staff to resist Executive Board pressure to support a country with IMF lending, either when inappropriate policies are being pursued (for example, the creation of the quasi currency board) or when a country is insolvent (as Argentina was by 2003). Second, the Fund has sometimes found it just as hard as its members to take a stand against an inappropriate fixed-exchange-rate regime. Third, the absence of any international standstill process or debt restructuring mechanism makes it difficult and time consuming to reconstruct a financial system and to reach a balanced solution with creditors once a crisis has occurred. The Taylor doctrine has not worked out wholly as planned. Fourth, once damaged, the quality of the policy dialogue between the Fund and its members is difficult to restore. Since the crisis, Argentina's policies have appeared unsustainable: Argentina has contrived to keep its exchange rate at a level at which its exports seem to be excessively competitive, while relying heavily on high international primary commodity prices to sustain its balance of payments. These policies do not seem consistent with the world envisaged in the second amendment of the Fund's Articles, a world in which the Fund exercises firm surveillance over member countries' policies in its role as steward of the international financial system.

Conclusions

The capital account crises of the 1990s and 2000s represent a new chapter in the Fund's history: they mark a distinct shift from the Fund's previous bread-and-butter work of dealing with current account crises. These capital account crises created new challenges and strains on the Fund – some of which it responded to well, some less so.

On *crisis prevention* the Fund has learned much. After the Mexican crisis it promoted regulatory reform, increased transparency, and better monitoring in emerging market economies. The Fund's Articles prevent it from pronouncing on countries' particular choice of exchange-rate regimes. But in its policy advice the Fund has made clear that the trilogy of floating exchange rates, carefully sequenced liberalization of capital accounts and financial systems, and inflation targeting can work well (Blejer et al. 2001; Corden 2002; Batini et al. 2005); by contrast, the Fund has given clear advice about the difficulties faced by fixed exchange-rate regimes. The Fund has also attempted to reinvent itself as a lender of 'first resort' through the creation of contingent or 'pre-approved' lending facilities aimed at crisis prevention. These lending windows would provide members with an added incentive to pursue sound policies and a signalling framework under which they could commit to these policies. But the Fund's first effort in this direction – 1999's Contingent Credit Lines (CCL) – expired in 2003 after four years without use, owing to somewhat stringent qualification criteria, less than full automaticity in disbursements, and concerns amongst members that a request for a CCL might send a negative signal to capital markets. New effort was invested in the design of such an instrument, initially called the Reserve Augmentation Line (RAL), during 2006–07.

On *crisis management*, much work has been done to understand better how to construct, balance and sequence macroeconomic policy

restraint at a time of crisis. The Fund has developed a detailed debt sustainability framework and complemented its traditional analysis of financial flows with a 'balance sheet approach' to analysing stock imbalances, so as to enable it to understand the financial vulnerabilities of countries. This tool was designed to help Fund staff draw a clearer distinction between liquidity crises and solvency cases. (On this see Irwin and Vines 2005; Cohen and Portes 2004; Portes 2004.) But from the early 1980s onward, the three generations of crises outlined above also threw into sharp relief the problem of moral hazard arising from IMF lending. The need to balance better debtor moral hazard and creditor moral hazard became one of the key challenges facing the Fund in the design of its lending facilities and its accompanying policy responses to crises. This article has highlighted the manner in which the Fund has occasionally oscillated between favouring creditor interests and favouring debtor interests, in an attempt to balance these interests in an acceptable way.

The Fund's experience with crisis management in the 1990s revealed difficulties with Fund conditionality. By then the conditions attached to Fund loans had grown far beyond what had earlier been thought necessary to ensure adequate macroeconomic adjustment, and came to include substantial structural conditionalities. Some of these concerned macroeconomic issues of proper concern to the Fund. But there was also an explicit concern with a range of microeconomic reform issues, and, even more broadly, with poverty-reduction questions. Many observers, including Arriazu et al. (1999), IFIAC (2000) and Williamson (2000), have questioned the wisdom of this policy creep, although it should be said that, in some cases (for example, poverty reduction), the spread of IMF conditionality reflected the concerns of member countries rather than an attempt by the Fund to expand its mandate. Following member country dissatisfaction with the comprehensive conditionalities included in their programmes (Indonesia's programmes in the late 1990s are particularly relevant cases), there has been much work at the IMF since 2000 on streamlining conditionality, and on pulling back from a range of concerns about structural issues that are not deemed 'macro critical'. This led to a careful restatement during 2002 of the principles governing the IMF's design and implementation of conditionality, with a view to ensuring that the conditions attached to IMF lending focus only on policies essential to the macroeconomic viability of Fund-supported programmes. (See IMF 2002a; Boughton and Mourmouras 2004.)

At the time of the preparation of this article (2007) there was a lull in the frequency of crises, and a significant decline in the volume of Fund lending. The Asian, Russian and Argentinean borrowings which originated in the crises described above had all been repaid. There is a striking parallel here with the end of the 1980s, when the Fund's stock of outstanding loans to emerging markets was also quite modest. At that time, the Latin American arrangements that had originated in the crisis years 1980–83 had been repaid. But, just as then, risks remain; the international community must remain engaged in the task of ensuring that the Fund is prepared to respond to and manage crises when they occur.

Dissatisfaction with the Fund's crisis management in the 1990s and early 2000s cast a long shadow over the Fund's relations with many emerging-market economies, which may have some consequences. A number of East Asian countries, over the ten years following the East Asian crisis, accumulated in excess of a trillion US dollars of reserves. This massive reserve accumulation reflected a persistent excess of saving over investment across these economies, which may, at least in part, represent a conscious choice to amass reserves as a form of self-insurance against future crises. These countries went about a pooling of some of these reserves into a common fund, a process which began in 2000 when ASEAN, Japan, China and the Republic of Korea agreed to set up a bilateral currency swap scheme known as the Chiang Mai Initiative. There were some suggestions that this might one day form the basis of an Asian regional alternative to the IMF that would be designed to help these countries to co-insure and spread risks. But taking this step would require difficult decisions by these countries in order to make surveillance between the pool's members effective and enforceable.

And such a common pool of reserves might also create its own form of moral hazard if it were to encourage countries to take excessive risks with foreign borrowing.

The IMF and Low-Income Countries

Until the mid-1970s, the Fund's work in its role as coordinator and monitor of the international monetary system was concerned mainly with monetary, exchange-rate and trade issues. To the extent that the IMF also functioned as a credit union for countries in balance of payments difficulties, its lending focused on the provision of short-term, self-liquidating loans to buttress central banks through temporary balance of payments difficulties. The Fund's cornerstone principle of equal treatment of member countries dictated that finance to low-income countries was provided largely under stand-by arrangements on the same terms as those approved for emerging markets and industrialized countries. The oil crises of the 1970s, however, made it increasingly clear that intractable structural issues in many low-income countries needed to be tackled if balance of payments difficulties were to be addressed. As a result, the 1970s saw a lengthening of the average maturity of stand-by arrangements in both emerging markets and low-income countries, accompanied by the advent of lending on concessional terms, with lower interest rates, to low-income countries. This created some tension between the Fund's essentially monetary character and its deepening role in the provision of longer-term resources in support of broad macroeconomic adjustment in developing countries.

In order to provide member countries with more breathing room to enact structural economic reforms, the Fund created a series of new lending instruments from the mid-1970s onward. The first amongst these, the Extended Financing Facility (EFF), provided greater financing and longer maturities than traditional stand-by arrangements, but its terms were not concessional. The Fund's Articles of Agreement did not provide for the use of IMF resources for concessional lending to a subset of the Fund's membership, and the EFF's market-linked interest rates were identical to those of other Fund arrangements. An EFF did, however, typically carry more stringent conditionality than a stand-by arrangement in response to concerns that the EFF's greater financing implied a need for greater adjustment.

The obstacle to financing concessional lending posed by the Fund's Articles was overcome in the 1970s by the solicitation of donor funds and the sale of a portion of the IMF's gold. Concessional IMF lending began under the 1975 Oil Facility Subsidy Account, in which contributions from 25 countries were used to reduce the interest cost of borrowing from a Fund facility set up to assist countries deemed to have been most severely affected by the sudden rise in oil prices. In the following year, the IMF created a Trust Fund for all low-income countries out of profits from the sale of a portion of the Fund's stock of gold. The Trust Fund offered long-term low-interest loans to low-income countries from 1976 until its resources were fully committed in 1981. Borrowing under the Trust Fund was similar to financing under the first credit tranche: in order to obtain financing, low-income countries had only to demonstrate a balance-of-payments need and explain the efforts they were taking to reduce it.

These new financing windows provided concessional loans to developing countries, but it was feared that the weak conditionality attached to these loans did not induce sufficient adjustment (Boughton 2001). In the early to mid-1980s prices for many primary commodities collapsed, and several developing countries faced new external balance of payments challenges. The Fund moved to reinvigorate its concessional lending by using the repayments of Trust Fund loans to finance a new round of concessional credit under what, in 1986, came to be known as the Structural Adjustment Facility (SAF). The SAF marked a determined attempt by the Fund to integrate concessionality with conditionality. In part, this twinning of concessionality with conditionality allowed the Fund to lobby for new donor loans and grants, which expanded the SAF some threefold into the Enhanced SAF (ESAF) in 1987.

Boughton (2001) contends that the ESAF became one of the IMF's great success stories, as

it allowed the Fund to send billions of dollars to the world's poorest counties on concessional terms with longer maturities than was possible under previous IMF facilities. (See also Tarp 1993.) The ESAF also had a catalytic effect on lending from other official creditors, and IMF collaboration with the World Bank and the regional development banks, as well as with, inter alia, the UN, UNICEF, UNDP and bilateral donors, all appeared to improve under the ESAF process (Boughton 2001). In addition, IMF technical assistance to many developing countries on monetary, fiscal, and trade policy, as well as debt management, also expanded substantially in order to help countries achieve their programme commitments. This increase in technical assistance has been very valuable.

Despite these gains, and even although the ESAF was technically distinct from the Fund's general resources, some critics have charged that the ESAF marked an unfortunate departure from the Fund's monetary focus. Others have questioned the strict conditionality on adjustment agreed under ESAF-supported programmes, especially because some of the structural conditions have appeared to intrude on the traditional territory of the World Bank. In reply it might be said that this has happened partly because the Bank has not proved capable of devising appropriate macroeconomic conditions for its own loans. (See Gilbert and Vines 2000.)

Despite the Fund's efforts – both to revive its concessional lending in 1986 and 1987 and to increase its accompanying technical assistance – it was clear by 1988 that many low-income countries would find it impossible to grow without debt relief. Under the auspices of the Paris Club of bilateral creditors, a series of progressively more concessional refinancing terms for bilateral debts were agreed from 1988 onward, for both emerging market, and relatively poor, indebted countries. Nevertheless, even with this bilateral debt relief, many low-income countries had trouble meeting the payment obligations on their standby arrangements and EFFs. But the absence of a serious lobby of private creditors (most low-income countries' external debt was owed to the Paris Club and other public creditors) may have delayed efforts to find a comprehensive solution to the debt problems of developing countries until the late-1990s.

By the 1990s, the Fund's engagement in low-income countries had become the target of a rising chorus of concern. Some civil society organizations and academics, as well as some low-income governments themselves, contended that IMF conditionality and programme design in low-income countries tended to prioritize adjustment over poverty reduction, growth, and income distribution concerns. This criticism is summarized by Easterly (2005). It arose despite the fact that the Fund has been helping to produce, in many low-income countries, a marked stabilization in macroeconomic indicators, and in some cases the beginning of sustained periods of growth. In response to critics' concerns, and in a further step in the evolution of Fund lending, IMF Managing Director Michel Camdessus advocated in the mid-1990s a fresh model of engagement with low-income countries in which there would be a renewed role for the Fund in reducing global poverty and in promoting high-quality growth in developing countries.

This new strategy featured three main elements. First, along with bilateral donors and other international financial institutions, the Fund recognized that catalysing growth in low-income countries would require more profound debt relief, including treatment of previously unrescheduled multilateral concessional debt. The 1996 Heavily Indebted Poor Countries' (HIPC) Initiative represented the concerted efforts of the international community to address the external debt overhang in poor countries; the Initiative was later enhanced in 1999 to provide deeper and faster debt reduction. The HIPC Initiative was novel, particularly in that debt relief was explicitly tied to plans to spend debt-service savings on poverty-alleviating social expenditure. From 1999, these plans were articulated in a country-based Poverty Reduction Strategy Paper (PRSP). This approach, initiated by the Fund in conjunction with the World Bank, formed the second prong of the Fund's renewed engagement with low-income countries. The PRSP approach aimed to provide a clear country-owned link between national policy frameworks, donor

support, and development outcomes. The PRSP approach also dovetailed neatly with the United Nations' Millennium Development Goals (MDGs). These goals were articulated at the UN Millennium Summit in 2000 and were centred on halving global poverty by 2015. The PRSPs were also intended to form the basis of the targets and policy conditions in programmes supported by the IMF's Poverty Reduction and Growth Facility (PRGF). This was the successor in 1999 to the ESAF and formed the third element of the Fund's new approach to low-income countries.

The results of these initiatives by the early 21st century were mixed. Reviews of the PRGF by IMF staff in 2002 (IMF 2002b) and by the IMF's IEO in 2004 (IMF 2004b) found that PRGF-supported programs had become more accommodating to higher public expenditure, in particular pro-poor spending. Nevertheless, a review of PRGF programme design by the IMF Executive Board in September 2005 (IMF 2005c) found that per capita income and growth rates remained low despite some improvements in a range of macroeconomic indicators. More recently, the IEO found in its evaluation of Fund engagement in sub-Saharan Africa (IMF 2007b) that the PRGF and PRSP approaches had not had a significant positive effect on catalysing new aid flows. This is despite the fact that commitments to increase such flows were made in 2002 under the 'Monterrey Consensus' and at the Gleneagles G8 summit in 2005. The IMF's Spring 2007 *Regional Economic Outlook* noted, however, that Sub- Saharan Africa's growth performance since 2004 had been the best in more than three decades (IMF 2007d). In sum, the impact of the PRGF and PRSP on aid and spending in low-income countries remained inconclusive, but their growth effects appeared increasingly positive by 2007.

The advent of the HIPC Initiative, the PRSP and the PRGF together intertwined the work of the IMF and World Bank in developing countries to an unprecedented extent. The Multilateral Debt Relief Initiative (MDRI) agreed at the Gleneagles G8 Summit in 2005, and which provided a framework for the write-off of nearly all remaining HIPC-country debts to the IMF, World Bank and African Development Bank, represented a major step forward in this collaboration. While the MDRI drew a welcome line under the multilateral debt relief process, it left several questions about the next phase of IMF and World Bank support for low-income countries unanswered. Having written off so much concessional debt, the MDRI implied that future multilateral support for low-income countries should be provided only as grants, not loans. The source of financing for such grants remained unclear. And in some cases, financing, whether by grants or loans, may not be the most crucial contribution that the international financial institutions could make to development. The Fund's 2005 Policy Support Instrument (PSI), essentially a 'no money' programme, acknowledged that Fund macroeconomic advice, rather than short-term balance of payments financing, might be a valuable channel of support for developing countries. These matters have been complicated by the growth of 'South–South' flows in development assistance from new donors such as China and Brazil. These flows have raised doubts about the future necessity of concessional financing from the Bretton Woods institutions. But they have also called into question the conditionality that comes attached to IMF and World Bank money. Such financing from non-traditional donors could also complicate future debt restructurings, should they prove necessary, since most new donors have not been members of the Paris Club.

Throughout this section we have noted the latent tension between the Fund's monetary character and its long-term support for low-income countries. This tension is heightened by the intertwining of the work of the Fund and the World Bank, which we have just reviewed. The report of the external review committee on Bank–Fund collaboration (IMF 2007c) provided some suggestions on strengthening Bank–Fund collaboration, while reducing overlap between the two institutions.

The Future of the IMF: Next Steps

In mid-2004 the Fund's Managing Director, Rodrigo de Rato, launched a review of the role of IMF in light of the challenges posed by a

changing and increasingly complex global economic system. Stemming from this review, De Rato presented the aforementioned Medium-Term Strategy for the Fund (IMF 2005b) to the World Bank–IMF Annual Meetings in September 2005, and shortly thereafter followed up with a plan for the Strategy's implementation (IMF 2006b). The plan focused on specific proposals to ensure that the Fund:

- Provides more effective surveillance and better monitoring of policies in advanced economies, with a renewed emphasis on exchange rates;
- Provides better monitoring of emerging markets economies, re-explores financing mechanisms to help prevent crises, and reconsiders issues regarding capital account liberalization;
- Enhances the role of IMF in low-income countries, and sharpens its focus;
- Reforms IMF governance, particularly country representation; and
- Restructures the IMF's own budget, including by broadening the Fund's income base, and its management practices.

The plan also expressed an intention to expand the role of the IMF as a provider of technical assistance and training, while improving Fund communications and transparency to ensure that the Fund would play a more central role in global policy debates.

The Fund's Medium-Term Strategy is a clear response to the three dominant tasks it has assumed following the collapse of the Bretton Woods system of fixed exchange rates in 1971, tasks which we have reviewed in Sections 3–5 of this article. But if the Fund is to be able to act effectively in relation to these tasks it will need to have: (i) a better system of governance; (ii) a more secure and robust source of income so that it can cover its operating expenses; and (iii) a larger stock of resources to lend for crisis prevention and resolution. We conclude this article by briefly discussing these three issues. (See also Lane 2006.)

Governance

The first subsection of the Fund's Articles of Agreement made clear that its founding purpose was 'to promote international monetary cooperation through a permanent institution which provides the machinery for consultation and collaboration on international monetary problems'. At the time of the Fund's creation, most countries stood a reasonable chance of alternating between being a creditor to and borrower from the Fund over time. Since then, the ranks of creditors and borrowers have diverged as industrial countries have stopped using IMF financing, a role which has instead been filled by emerging market economies and low-income countries. A number of reformers such as Woods (2006) argue that the Fund's capacity to facilitate solutions to international monetary problems depends on the Fund's decision-making structure being made more reflective of the interests and voices of the emerging markets and developing countries which borrow from it, and which see their public policy frameworks at least partly determined by Fund conditionality. The demand for such reform is bolstered by the fact that the relative distribution of quotas, which determine the voting power in the Fund, has become separated from the relative economic (and political) weights of many emerging markets in the global economy. In addition, the relative power of basic votes, which were intended to provide some measure of fairness to poorer countries, has been substantially eroded relative to the contribution of quotas to voting weights at the Executive Board. The ad hoc provision of increased quota shares to China, Korea, Mexico, and Turkey in 2006 under the Fund's Medium-Term Strategy was a first step toward realigning voting power in the Fund with emerging markets' growing share of the world economy; further steps will be more difficult since increased voting shares for some countries will inevitably mean painful decisions to reduce the shares of others. It may, however, be possible for countries to change the way in which the 24 chairs on the IMF's Executive Board are allocated in order to compensate partly for changes in relative voting shares.

Changing the Fund's voting structure would not in and of itself alter the way in which the Fund operates, suddenly making it better able to deliver on the objectives set out in its 2005

Medium-Term Strategy. De Gregorio et al. (1999); King (2006a, p. 12); Dodge (2006a, b) and Kenen (2006) have all argued, however, that parallel changes in the Fund's governance arrangements might help the Fund in its push towards these objectives.

One proposal would put the responsibility for the delivery of improved policies more firmly in the hands of the management of the IMF. Up to 2007, the Executive Board of the Fund had involved itself in day-to-day reviews of Article IV reports, approved all lending decisions, and reviewed the design of the Fund's lending programmes. Stepping back from this activity would enable Directors to pay proportionately more attention to strategic issues. That would move the governance structure of the Fund closer to the relationship between management and advisory boards that one sees in the private sector, where non-executive directors bring dispassionate external views to broad questions of corporate operations and strategy, and clearly delegate day-to-day operations to management.

Evolution in this direction could strengthen the accountability of the Managing Director and his Deputies. In one version of this type of arrangement, all of the Managing Director, the Deputy Managing Directors, and Department Directors would report on a regular basis to the Board, but Executive Directors would be more removed from many of the day-to-day decisions of the institution. Doing this could have an effect – even if only implicit or indirect – on the Fund's ability to function better in its pursuit of more dispassionate surveillance. It might also lead to more effective crisis prevention and resolution through a careful balancing of debtor moral hazard and creditor moral hazard in Fund lending; and also to a clearer focus in the Fund's work with low-income countries.

A move to a non-resident Executive Board would draw a clearer line between the work of Directors and management. Such a move would leave the Managing Director in control of the execution of the Fund's work since the Executive Directors would give only part-time oversight and direction. Making this change would take the governance of the Fund closer to Keynes' original vision. (See King 2006a.) Directors would be the senior public servants that steer policy in their national capitals, and not, as in 2007, their proxies resident in Washington. In contrast with 1946, the ease of modern travel makes a non-resident Board, with meetings some six to eight times a year, entirely feasible. Any move in this direction would, however, need to ensure that the nexus of communication between capitals, which the Board currently provides, is preserved in some other way.

Income

In May 2006, the Managing Director established a committee (the 'Crockett Committee'), chaired by a former General Manager of the Bank for International Settlements, Andrew Crockett, to study options for sustainable long-term financing of the IMF. The Committee's report, released on 31 January 2007 (IMF 2007a), argued that the IMF's current funding model was unsustainable and that a more diversified income stream needed to be developed in order to guarantee the institution's financial future.

The IMF's revenue stream had been primarily based on income derived from its lending for crisis resolution (IMF 2007a, Annex 2, p. 2). This financing mechanism was not entirely appropriate, because, as Crockett said during the press briefing to launch the Committee's report, 'it's a concentrated income source … It's volatile, because when the Fund is lending a lot. . .it generates large resources. When the Fund is not lending, it doesn't generate resources.' In a low-lending environment, as existed in the early 21st century, the Fund's income model appeared untenable over the longer term; in the shorter term, it could also be inconsistent with sound incentives to minimize moral hazard in Fund lending.

The Committee considered some alternative sources of income for the Fund. In assessing these possibilities, the committee observed that the Fund's activities could be broken down into three types of functions that cut across the full membership of industrialized countries, emerging markets, and low-income economies: financial intermediation, the provision of global public

goods (for example, data, standards and codes, and combating terrorist financing), and the provision of bilateral services, in the form of capacity building and technical assistance.

The Committee concluded that revenue from Fund lending should be sufficient to cover its ongoing costs arising from financial intermediation. The Committee also noted that this income should not be used to cross-subsidize the provision of global public goods because (i) this income was too volatile for this purpose and (ii) cross-subsidization could cause IMF lending to become too expensive compared with private financing.

In order to ensure that the Fund could continue to provide its key global public goods, the Committee noted that the Fund could, like the United Nations, assess a periodic levy on member countries. The Committee did not, however, favour this source of income, as it 'would risk politicising the activities of the Fund' by making its work subject to regular financing calls. Nevertheless, the Committee did note that charging fees for some services might generate a small amount of additional revenue.

The Committee's core proposal concerned the creation of an endowment for the IMF that would provide a reliable income stream without relying on annual requests to member countries. The Committee suggested a further sale of IMF gold as a possible source of endowment funds. Such sales had been mooted at various points in the past for a variety of purposes; this was done to finance the establishment of the trust funds that underwrote the 1996 HIPC Initiative. But other plans for such sales have usually failed to gain enough support in the face of opposition from the United States and from gold-producing countries. To allay these fears, the Committee report suggested a 'balanced' approach, in which the Fund would also invest some of its quota resources in highly rated securities so that the burden of creating an investment endowment would not fall exclusively on the sale of gold.

As this article was being drafted, discussion was continuing on the exact form an endowment for the Fund could take. In meantime, the Fund had begun to invest some of its retained earnings from lending in investment grade securities in an effort to supplement its income.

Resources

The relative size of the Fund shrank markedly from the 1970s ownard in comparison with, inter alia, global reserves, international trade, financial flows, stocks of financial assets and world output. This decline in pecuniary stature has distorted some of the debates about the Fund's work, most notably on creditor and debtor moral hazard. Much of the debate over the implications of jumbo or 'exceptional access' arrangements in the 1990s (arrangements in which lending was equivalent to 300 per cent of quota or more) would be moot if regular quota increases had maintained the Fund's relative size in the global economy. Indeed, had the Fund grown through regularly scheduled quota increases, very few of the arrangements of the 1990s and 2000s would have been deemed at all exceptional. This suggests a simple yardstick for an appropriately-sized IMF: at any given time, the sum of the Fund's quotas should enable a risk-adjusted subset of its membership to borrow from the Fund on non-exceptional terms to finance their adjustment needs.

Accepting the validity of such a yardstick depends critically, however, on one's ultimate view of the role the IMF should play in the international system: trusted macroeconomic advisor, catalyst for private capital inflows and foreign assistance, or potential lender of last resort at time of crisis? To some extent the Fund played all of these roles at the turn of the 21st century, though its reduced relative size meant that the lender-of-last-resort function was credible only for its smaller members. The Fund staff, its shareholders, and those who care about the future of the multilateral system will need to decide which of these roles the IMF should continue to play.

See Also

- ▶ Bretton Woods System
- ▶ Currency Crises
- ▶ Development Economics

- ▶ Emerging Markets
- ▶ Keynes, John Maynard (1883–1946)
- ▶ Monetary Approach to the Balance of Payments
- ▶ World Bank

Bibliography

Agénor, R., M. Miller, D. Vines, and A. Weber (eds.). 1999. *The Asian financial crises: Causes, contagion, and consequences*. Cambridge: Cambridge University Press.

Arriazu, R., J. Crow, and N. Thygesen. 1999. *External evaluation of IMF surveillance*. Washington, DC: IMF. Online. Available at http://www.imf.org/external/pubs/ft/extev/surv/eval.pdf. Accessed 6 June 2007.

Batini, N., K. Kuttner, and D. Laxton. 2005. Does inflation targeting work in emerging markets? In *World economic outlook,* September. Washington, DC: IMF. Online. Available at http://www.imf.org/external/pubs/ft/weo/2005/02/pdf/chapter4.pdf. Accessed 6 June 2007.

Blejer, M., A.M. Leone, P. Raubanal, G., and Schwartz. 2001. Inflation targeting in the context of IMF-supported adjustment programs. Working Paper No. 01/31. Washington, DC: IMF.

Bluestein, P. 2001. *The Chastening: Inside the crisis that rocked the international financial system and humbled the IMF*. New York: Public Affairs.

Boorman, J., T. Lane, M. Schulze-Ghattas, A. Bulir, et al. 2000. Managing financial crises: The experience in east Asia. Working Paper No. 00/107. Washington, DC: IMF.

Bordo, M.D. 1993. The Bretton Woods international monetary system: An historical overview. In *A retrospective on the Bretton Woods system, lessons for international monetary reform*, ed. M. Bordo and B. Eichengreen. Chicago/London: University of Chicago Press.

Boughton, J. 2001. *Silent revolution: The international monetary fund 1979–1989*. Washington, DC: IMF.

Boughton, J. 2004. The IMF and the force of history: Ten events and ten ideas that have shaped the institution. Working Paper, No. 04/75. Washington: IMF.

Boughton, J., and A. Mourmouras. 2004. Whose programme is it? Policy ownership and conditional lending. In *The IMF and its critics: Reform of global financial architecture*, ed. D. Vines and C. Gilbert. Cambridge: Cambridge University Press.

Calvo, G., and M. Goldstein. 1996. What role for the official sector? In *Private capital flows to emerging markets after the Mexican crisis*, ed. G. Calvo, M. Goldstein, and E. Hochreiter. Washington, DC: Institute for International Economics.

Calvo, G., and E. Mendoza. 1996. Reflections on Mexico's balance of payment crisis: A chronicle of a death foretold. *Journal of International Economics* 41: 235–264.

Cohen, D., and R. Portes. 2004. Towards a lender of first resort. Discussion Paper No. 4615. London: CEPR.

Corbett, J., and D. Vines. 1999a. Asian currency and financial crises: Lessons from vulnerability, crisis, and collapse. *World Economy* 22: 155–77.

Corbett, J., and D. Vines. 1999b. The Asian crisis: Lessons from the collapse of financial systems, exchange rates, and macroeconomic policy. In Agénor et al. (1999).

Corbett, J., G. Irwin, and D. Vines. 1999. From Asian miracle to Asian crisis: Why vulnerability, why collapse? In *Capital flows and the international financial system*, ed. D. Gruen and L. Gower. Sydney: Reserve Bank of Australia. Repr. in Irwin and Vines (2001).

Corden, W.M. 1983a. Is there an important role for an international reserve asset such as the SDR? In *International money and credit: The policy roles*, ed. G.M. von Furstenberg. Washington, DC: IMF.

Corden, W.M. 1983b. The logic of the international monetary non-system. In *Reflections on a troubled world economy*, ed. F. Machlup, G. Fels, and H. Muller-Groeling. London: Macmillan.

Corden, W.M. 1990. Macroeconomic adjustment in developing countries. In *Public policy and economic development*, ed. M. Scott and D. Lal. Oxford: Clarendon Press.

Corden, W.M. 1993. Why did the Bretton Woods system break down? In *A Retrospective on the Bretton Woods system*, ed. M. Bordo and B. Eichengreen. Chicago: University of Chicago Press.

Corden, W.M. 1999. *The Asian crisis: Is there a way out?* Singapore: Institute of South Asian Studies.

Corden, W.M. 2002. *Too sensational: On the choice of exchange rate regimes*. Cambridge, MA: MIT Press.

Corden, W.M. 2007. Those current account imbalances: A sceptical view. *World Economy* 30: 363–382.

Dam, K. 1982. *The rules of the game*. Chicago: University of Chicago Press.

De Gregorio, J., B. Eichengreen, T. Ito, and C. Wyplosz. 1999. *An independent and accountable IMF*. (Geneva Reports on the World Economy, No. 1). London: CEPR.

de Vries, M.G. 1976. *The international monetary fund 1966–71: The system under stress*. Washington, DC: IMF.

de Vries, M.G. 1985. *The international monetary fund 1972–78: Cooperation on trial*. Washington, DC: IMF.

de Vries, M.G. 1987. *Balance of payments adjustment, 1945 to 1986: The IMF experience*. Washington, DC: IMF.

Dodge, D. 2006a. Global imbalances: Why worry? What to do? Speech made to the New York association for business economics, 29 March. *BIS Review,* 4 April. Online. Available at http://www.bis.org/review/r060404a.pdf. Accessed 19 June 2007.

Dodge, D. 2006b. The evolving international monetary order and the need for an evolving IMF. Speech given at the Woodrow Wilson School of public and international affairs, 30 March. *BIS Review,* 5 April. Online. Available at http://www.bis.org/review/r060405a.pdf. Accessed 19 June 2007.

I

Dooley, M., F. Landau, and P. Garber. 2003. An essay on the revived Bretton Woods system. Working Paper No. 9971. Cambridge, MA: NBER.

Easterly, W. 2005. What did structural adjustment adjust? The association of policies and growth with repeated IMF and World Bank adjustment loans. *Journal of Development Economics* 76: 1–22.

Edwards, S. 1989. The international monetary fund and the developing countries: A critical evaluation. In *IMF policy advice, market volatility, commodity price rule, and other essays*, ed. K. Brunner and A.H. Meltzer. Carnegie-Rochester Conference Series on Public Policy. Amsterdam: North-Holland.

Eichengreen, B. 1992. *Golden fetters: The gold standard and the great depression 1919–1939*. Oxford: Oxford University Press.

Eichengreen, B. (ed.). 1995. *Europe's postwar growth*. New York: Cambridge University Press.

Eichengreen, B. 2004. The dollar and the new Bretton Woods system. Henry Thornton Lecture, delivered at the Cass School of Business. London, 15 December.

Eichengreen, B. 2007. *The European economy since 1945: Coordinated capitalism and beyond*. Princeton: Princeton University Press.

Eichengreen, B., and R. Portes. 1995. *Crisis, what crisis? Orderly workouts for sovereign debtors*. London: CEPR.

Eichengreen, B., and Y.C. Park. 2006. Global imbalances and emerging markets. Paper presented to the Asia-Europe Economic Forum Conference in Beijing on 13 July, 3, 7–68.

Finch, C.D. 1989. *The IMF: The record and the prospects*, Essays in international finance no. 175. Princeton: International Finance Section, Department of Economics, Princeton University.

Fischer, S. 1999. On the need for an international lender of last resort. *Journal of Economic Perspectives* 13(4): 85–104.

Fischer, S. 2004. *IMF essays from a time of crisis: The international financial system, stabilization, and development*. Cambridge, MA: MIT Press.

Gardner, R. 1956. *Sterling dollar diplomacy.* Oxford: Oxford University Press.

Gilbert, C., and D. Vines. 2004. The IMF and international financial architecture: Liquidity and solvency. In Vines and Gilbert.

Gilbert, C., and D. Vines (eds.). 2000. *The World Bank: Structure and policies*. Cambridge: Cambridge University Press.

Goldstein, M. 2006. Currency manipulation and enforcing the rules of the international monetary system. In *Reforming the IMF for the 21st Century,* ed. E. Truman. (Special Report No. 19). Washington, DC: Institute for International Economics. Online. Available at http://www.iie.com/publications/chapters_preview/3870/05iie3870.pdf. Accessed 6 June 2007.

Gwin, C., J. Sachs, P. Kenen, R. Feinberg, and J. Nelson. 1990. *The international monetary fund in a multipolar World: Pulling together.* US–Third World Policy Perspectives No. 13, Overseas Development Council. New Brunswick: Transaction Books.

Horsefield, J.K. 1969. *The international monetary fund 1945–65: Twenty years of international monetary cooperation*, 3 vols. Washington, DC: IMF.

IFIAC (International Financial Institution Advisory Commission). 2000. *Report of the international financial institution advisory commission*. Washington, DC: IFIAC, US Congress.

IMF. 2002a. Guidelines on conditionality. Prepared by the Legal and Policy Development and Review Departments, 25 September. Online. Available at http://www.imf.org. Accessed 30 May 2007.

IMF. 2002b. Review of the poverty reduction and growth facility (PRGF) and the poverty reduction strategy paper (PRSP) Approach, 15 March. Online. Available at http://www.imf.org. Accessed 30 May 2007.

IMF. 2003. *The IMF and recent capital account crises: Indonesia, Korea, Brazil*, Report of the Independent Evaluation Office of the IMF. Washington, DC: IMF. Online. Available at http://www.imf.org. Accessed 6 June 2007.

IMF. 2004a. Policy formulation, analytical frameworks and policy design. Prepared by the Policy Development and Review Department. Online. Available at http://www.imf.org. Accessed 30 May 2007.

IMF. 2004b. *Evaluation of the IMF's role in poverty reduction strategy papers and the poverty reduction and growth facility.* Report of the Independent Evaluation Office of the IMF. Washington, DC: IMF. Online. Available at http://www.imf.org. Accessed 6 June 2007.

IMF. 2005a. *Argentina: 2005 Article IV consultation – Staff report.* (IMF Country Report No. 05/236). Washington, DC: IMF.

IMF. 2005b. The managing director's report on the fund's medium-term strategy, 15 September. Online. Available at http://www.imf.org. Accessed 30 May 2007.

IMF. 2005c. Review of PRGF program design – Overview. 8 August. Policy Development and Review Department. Online. Available at http://www.imf.org. Accessed 6 June 2007.

IMF. 2006a. *Multilateral Surveillance.* Report of the Independent Evaluation Office of the IMF. Washington, DC: IMF. Online. Available at http://www.imf.org. Accessed 6 June 2007.

IMF. 2006b. The Managing Director's report on implementation of the fund's medium-term strategy. 5 April. Online. Available at http://www.imf.org. Accessed 30 May 2007.

IMF. 2007a. Committee to study sustainable long-term funding of the IMF: Final report. Online. Available at http://www.imf.org. Accessed 6 June 2007.

IMF. 2007b. *The IMF and aid to sub-saharan Africa*. Report of the Independent Evaluation Office of the IMF. Washington, DC: IMF. Available at http://www.imf.org. Accessed 6 June 2007.

IMF. 2007c. Final report of the external review committee on bank–fund collaboration. February. Washington, DC: IMF. Online. Available at http://www.imf.org. Accessed 6 June 2007.

IMF. 2007d. *Regional economic outlook: Sub-Saharan Africa.* Washington, DC: IMF. Online. Available at http://www.imf.org/external/pubs/ft/reo/2007/AFR/ENG/sreo0407.htm. Accessed 6 June 2007.

IMF (International Monetary Fund). 1987. *Theoretical aspects of the design of fund-supported adjustment programs*, Occasional Paper No. 55. Washington, DC: IMF.

Irwin, G., and D. Vines (eds.). 2001. *Financial market integration and international capital flows*. Cheltenham: Edward Elgar.

Irwin, G., and D. Vines. 2003. Government guarantees, investment, and vulnerability to financial crises. *Review of International Economics* 11: 860–874.

Irwin, G., and D. Vines. 2005. Policies for the resolution of international financial crises: How to avoid moral hazard. *International Journal of Money and Finance* 10: 233–250.

Joshi, V., T. Lane, and D. Vines. 2006. *The US, East Asia and Europe: How to achieve an orderly resolution of global imbalances*. Mimeo: Oxford University.

Kenen, P. 2006. *Comments on the address of the Managing Director of the IMF*. Washington, DC: Institute for International Economics. Online. Available at http://www.iie.com. Accessed 30 May 2007.

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

Keynes, J.M. 1971–1988. *The collected writings of J.M. Keynes,* ed. E. Johnson and D. Moggridge. London: Macmillan.

Killick, T. 1986. *The quest for economic stabilisation: The IMF and the third world*. Aldershot: Gower.

King, M. 2005. The international monetary system. Remarks at Advancing Enterprise 2005 Conference, London, 4 February. Online. Available at http://www.bankofengland.co.uk. Accessed 30 May 2007.

King, M. 2006a. Reform of the international monetary fund. Speech given to the Indian Council for Research on International Economic Relations. New Delhi, 20 February. Online. Available at http://www.bankofengland.co.uk. Accessed 30 May 2007.

King, M. 2006b. Through the looking glass: Reform of the international institutions. Inaugural International Distinguished Lecture, Melbourne Centre for Financial Studies, Australia, 21 December. Online. Available at http://www.bankofengland.co.uk. Accessed 30 May 2007.

Krueger, A. 2002. A new approach to sovereign debt restructuring. Address given at the Indian Council for Research on International Economic Relations, Delhi, 20 December. Online. Available at http://www.imf.org. Accessed 6 June 2007.

Krugman, P. 1999. Balance sheets, the transfer problem, and financial crises. *International Tax and Public Finance* 6: 459–472.

Kumar, M. 2006. Economic policy cooperation and coordination: Analytical issues and historical experience. Mimeo, Fiscal Affairs Department, IMF.

Lane, T. 2006. Tensions in the role of the IMF and directions for reform. *World Economics* 6(2): 47–66.

Little, I.M.D. 1993. *Macroeconomic analysis and the developing countries, 1970–1990*. Occasional Paper No. 41. San Francisco: International Centre for Economic Growth.

Little, I.M.D., R. Cooper, W.M. Corden, and S. Rajapatirana. 1993. *Boom, crisis and adjustment: The macroeconomic experience of developing countries*. Oxford: Oxford University Press.

Matthews, R.C.O., and A. Bowen. 1988. Keynesian and other explanations of postwar macroeconomic trends. In *Keynes and economic policy*, ed. W.A. Eltis and P.J.N. Sinclair. London: NEDO.

Matthews, R.C.O., C.H. Feinstein, and J.C. Odling-Smee. 1982. *British economic growth 1956–73*. Oxford: Clarendon Press.

McKibbin, W. 1997. Empirical evidence on international economic policy coordination. In *Handbook of comparative economic policies volume 5: Macroeconomic policy in open economies*, ed. M. Fratianni, D. Salvatore, and J. Von Hagen. London: Greenwood Press.

Meyer, L., B. Doyle, J. Gagnon, and D. Henderson. 2004. International coordination of macroeconomic policies: Still alive in the new millennium? In *The IMF and its critics: Reform of global financial architecture*, ed. D. Vines and C. Gilbert. Cambridge: Cambridge University Press.

Moggridge, D. 1986. Keynes and the international monetary system 1909–1946. In *International monetary problems and supply side economics: Essays in honour of lorie tarshis*, ed. J.S. Cohen and G.C. Harcourt. London: Macmillan.

Mussa, M., and M. Savastano. 1999. The IMF approach to economic stabilization. In *NBER macroeconomics annual 1999*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Odling-Smee, J. 2004. The IMF and Russia in the 1990s. Working Paper No. 04/155. Washington, DC: IMF.

Polak, J. 1957. Monetary analysis of income formation and payments problems. *IMF Staff Papers* 6: 1–50.

Portes, R. 2004. *Resolution of sovereign debt crises: The new old framework*, Discussion paper No. 4717. London: CEPR.

Roubini, N., and B. Setser. 2005. Will the Bretton Woods 2 regime unravel soon? The risk of a hard landing in 2005–2006. Unpublished manuscript. Online. Available at http://pages.stern.nyu.edu/∼nroubini/papers/BW2-Unraveling-Roubini-Setser.pdf. Accessed 19 June 2007.

Sachs, J. 1995. Do we need an international lender of last resort? F.D. Graham Lecture at Princeton University, 20 April.

Skidelsky, R. 2000. *John Maynard Keynes: Fighting for Britain, 1937–1946*. London: Macmillan.

I

Stiglitz, J. 2002. *Globalization and its discontents*. New York: W.W. Norton & Company.
Tarp, F. 1993. *Stabilization and structural adjustment: Macroeconomic frameworks for analysing the crisis in Sub-Saharan Africa*. London: Routledge.
Temin, P. 2002. The golden age of European growth reconsidered. *European Review of Economic History* 6: 3–22.
van Dormael, A. 1978. *Bretton Woods: Birth of a monetary system*. London: Macmillan.
Vines, D. 2003. John Maynard Keynes 1937–1946: The creation of international macroeconomics. *Economic Journal* 113: F338–F361.
Vines, D., and C. Gilbert (eds.). 2004. *The IMF and its critics: Reform of global financial architecture*. Cambridge: Cambridge University Press.
Williamson, J. 1977. *The failure of World monetary reform 1971–74*. New York: New York University Press.
Williamson, J. 1983a. Keynes and the international economic order. In *Keynes and the Modern World*, ed. G.D.N. Worswick and J. Trevithic. Cambridge: Cambridge University Press.
Williamson, J. (ed.). 1983b. *IMF Conditionality*. Washington, DC: Institute for International Economics.
Williamson, J. 2000. *The role of the IMF: A guide to the reports. Policy Brief 00–5*. Washington, DC: Institute for International Economics.
Williamson, J. 2006. The potential of international policy coordination. Paper presented at a seminar on implications for the IMF's role in surveillance and policy coordination, at a roundtable on international economic cooperation for a Balanced World Economy, Chongqing, China, 12–13 March. Online. Available at http://www.iie.com. Accessed 30 May 2007.
Wolf, M. 2005. Multilateral leadership can right the ship. *Financial Times,* 28 June.
Wolf, M. 2006. The world needs a tough and independent monetary fund. *Financial Times*, 22 February.
Woods, N. 2006. *The globalisers: The IMF, the World Bank and their borrowers*. Ithaca/London: Cornell University Press.

International Monetary Institutions

Stanley W. Black

Abstract

International monetary institutions are required to support payments arrangements between countries with different currencies and exchange rate arrangements. Reserve assets and adjustment and financing mechanisms are provided to assist markets in balancing conflicting objectives including economic growth and price stability, growing international trade and payments, and convertibility of currencies at reasonably stable exchange rates. The evolution of the Bretton Woods system has proceeded through floating exchange rates, increased capital mobility, financial crises, and various reform proposals. The development of regional monetary institutions has led to creation of the European Monetary Union and some steps towards increased Asian monetary cooperation.

Keywords

Asian Development Bank; Association of South East Asian Nations (ASEAN); Bretton Woods system; Capital controls; Competitive devaluation; Conditionality; Convertibility; European Central Bank; European Currency Unit (ECU); European Monetary System (EMS); European Monetary Union (EMU); European Payments Union; Exchange rate risk; Fixed exchange rates; Flexible exchange rates; Foreign exchange controls; General Arrangements to Borrow (IMF); Gold exchange standard; Gold standard; Inflation; International financial adjustment; International liquidity; International Monetary Fund; International monetary institutions; International reserves; Maastricht Treaty (EU); Monetary policy; Pegged exchange rates; Political risk; Reserve assets; Smithsonian Agreement (1971); Special drawing rights (SDRs); Stability and Growth Pact (EU); Tchebychev's inequality; Triffin, R.

JEL Classifications

F3

Domestic money is conceived of by society as a device to facilitate transactions in the marketplace, as a temporary store of value, and as a unit of account for contracts. Given the possibilities of fraud and counterfeiting, domestic monetary authorities have been established to regulate

the quality of the domestic monetary unit in most countries. Such regulations attempt to guarantee the interchangeability of the different media, such as currency and the deposits of different banks, as well as stability in the value of the monetary unit, under conditions of prosperity.

International monetary arrangements are required under conditions of international trade when residents of different countries must make payments to each other, and yet wish to hold most of their assets in terms of domestic currency. Such arrangements are designed to guarantee *convertibility* of assets denominated in different currencies, so that payments may be made independent of country of residence, thus facilitating a free and open trading system. International monetary *institutions* such as the International Monetary Fund are designed to support international monetary arrangements by enforcing rules of behaviour, assisting countries in difficulties, and encouraging good practices.

Alternative Exchange Rate Mechanisms

Under a *gold standard,* domestic residents and foreign residents may freely convert domestic currency into gold at a fixed rate of exchange. This type of convertibility was eliminated in the 1930s in favour of a gold *exchange* standard, which allowed only foreign monetary authorities to exchange domestic currency for gold. Gold convertibility of both types was ended as part of the Smithsonian Agreement of 1971 (see below).

Under a system of *pegged* exchange rates between different currencies, as established by the Bretton Woods system (see below), convertibility implies that domestic residents are free to obtain foreign currency at a *fixed* rate of exchange for the purchase of foreign goods and services, inclusive of normal trade credit. Likewise, foreign residents are free to sell domestic currency obtained by sale of goods and services or to use it for purchase of domestic goods and services, at the same fixed rate of exchange. This definition does not require free convertibility for capital account transactions (those arising from exchanges of financial assets only).

Under a system of floating or flexible exchange rates, convertibility still implies that both domestic and foreign residents may freely convert domestic and foreign currency at the same rate of exchange for current account transactions, but the exchange rate at which this may be done is determined on a daily basis by market transactions, rather than being guaranteed by the domestic monetary authorities of the respective countries.

In 2005, only 20 out of the 184 member countries of the International Monetary Fund (IMF) declined to accept the obligations to current account convertibility. But in a large number of countries various types of restrictions limited convertibility in some way or created differences in the exchange rates applying to exports and imports. Non-unified exchange rates lead to inefficient allocation of resources, as previously documented by Bhagwati (1978). For example, 70 countries required repatriation and surrender of proceeds of exports or invisible transactions, 57 countries had payments arrears of one kind or another, and 11 countries maintained either dual or multiple exchange rates for different types of transactions. With respect to capital account transactions, the situation is much more restrictive: 126 countries had controls on international transactions in capital market securities, and 143 countries maintained controls on direct investment flows.

Reserve Assets

In order to guarantee convertibility of the domestic currency into other convertible currencies, monetary authorities hold stocks of *reserve assets*, which are liquid assets held in readily accepted international media of exchange, such as dollars, euros, and a few other currencies. In addition, IMF member countries have access to unconditional borrowing rights to obtain additional reserve assets in the form of their reserve positions in the Fund and Special Drawing Rights. These, together with reserve asset holdings, make up *international liquidity.*

Since most international payments are handled by inter-bank transactions, banks have sought to

minimize transactions costs by channelling their foreign exchange transactions through one or more *vehicle* currencies, the pound sterling in earlier days, but more recently the US dollar and to some extent the euro. Because the dollar is so widely used in private exchange transactions, monetary authorities also find it convenient to operate in dollars to ensure the convertibility of their currencies.

Adjustment Mechanisms

The existence of different national currencies and the need to maintain convertibility of the different currencies lead to the concept of balance of payments adjustment mechanism. At a given exchange rate, as long as the amount of foreign exchange earned through exports of goods and services and capital inflows just pays for imports and capital outflows, no external imbalance exists. If international capital markets were perfect and if investors were risk neutral so that assets denominated in different currencies were perfect substitutes for one another in private portfolios, there would in practice be a single world interest rate for short-term borrowing. Then imbalances between foreign exchange earnings and payments could simply be *financed* by borrowing in the international capital market. There would be no real distinction between the convertibility characteristics of the official liabilities of different borrowers.

But, in fact, countries face very real limits on the amount of foreign currency they can borrow abroad in exchange for domestic currency because of *exchange rate* risk, which limits the willingness of risk-averse foreign lenders to acquire domestic currency assets. According to the doctrine of *original sin,* countries with a history of convertibility problems are unable to issue foreign debt in their own currency (Eichengreen and Hausmann 2005). The ability to repay foreign currency debt is dependent on balance of payments adjustment. *Political* risk involves the possibility that exchange controls may be imposed in the future, preventing the repayment of foreign currency debt on the promised terms. Thus it is desirable for countries to have access to a variety of adjustment mechanisms to eliminate external imbalances, as well as a variety of sources of official financing in the form of international liquidity. The primary mechanisms of balance of payments adjustment are through movements in exchange rates and adjustments of income and price levels via monetary and fiscal policies. The need for adjustment can be postponed by imposition of tariffs and subsidies, quantitative restrictions on current account or capital account transactions, or controls over the allocation of foreign exchange. But tariffs, quantitative restrictions, and exchange controls generally involve inefficiencies in the allocation of resources, including in the latter case loss of convertibility of the domestic currency. Changes in monetary and fiscal policies or exchange rates have their own costs in terms of domestic policy objectives forgone.

Financing

Thus, a mixture of adjustment policies and financing mechanisms is provided in a system of international monetary arrangements. *Official* financing is provided either by drawing on holdings of official reserve assets or by borrowing from international institutions. *Private* financing can be arranged by a monetary authority borrowing from foreign banks or the international bond market. Either provides the ability to postpone adjustment. The optimum mix of adjustment and financing for an individual country depends on the costs of the various alternatives. By setting the costs of these alternatives, international monetary arrangements influence the behaviour of the world economy.

A Model of Adjustment Versus Financing

In the theory of adjustment versus financing, a country is faced with random balance of payments deficits and surpluses, which it may either finance by drawing on reserve assets or adjust by one of the adjustment mechanisms mentioned above. In one branch of the theory, due to Heller (1966) and others, the cost of adjustment is assumed to be a

linear function of the size of the adjustment, so that any adjustments are postponed to the last minute, at which time full adjustment takes place. Alternatively, one may assume a nonlinear cost of adjustment, leading to a theory of partial adjustment. Kelly (1970) and Clark (1970) assume that the country's welfare function depends on the mean and variance of income, so that gradual adjustments are preferred. The analysis determines both the optimum level of reserve holdings, R^*, and the optimum rate of adjustment α to that level, according to the equation

$$\Delta R = \alpha[R^* - R_{-1}] + u, \qquad (1)$$

where u is normally distributed with mean zero and variance σ^2 and $R_{-1}$is the stock of reserves at the end of the previous period. This equation assumes that changes in the stock of reserves arise from both the random shocks in the balance of payments and the desired rate of adjustment to the optimal level of reserves. From Eq. (1) we find that the variance of reserve holdings decreases as the speed of adjustment α increases from zero to one.

Tchebychev's inequality then enables one to show that, for a given probability of not exhausting reserves and given opportunity cost r of holding reserves, the optimum reserve holding R^* decreases with increasing α. As α increases, the need for more frequent adjustments raises the variance of income. Therefore the speed of adjustment should be chosen such that the welfare loss from increased variance in income due to a small increase in α is just counterbalanced by the welfare saving due to holding slightly smaller reserves.

According to this theory, international monetary institutions will strongly affect the behaviour of national policies concerning balance of payments adjustment and acquisition of reserves. Specifically, international money institutions will determine the opportunity cost of holding reserves, the penalty attached to running out of reserves, and the availability of different types of adjustment policies. By influencing countries' balance of payments adjustment policies, international institutions will also influence their domestic policies, since there is a trade-off between internal and external objectives of policy.

The Role of Markets and Institutions

An optimal design for the international monetary system depends on balancing among a group of conflicting objectives: growth of real income and employment, stable prices, efficient allocation of resources, maintenance of convertibility of currencies, improving the distribution of income, and growth of world trade. The relevant trade-offs can be understood in the context of an economic model. According to the model of adjustment and financing outlined above, reductions in the opportunity cost of holding reserves will lead to increased reserve holdings, a reduction in the speed of adjustment to imbalances, increased use of financing, and a decline in the variability of income. The slowdown in the speed of adjustment implies a change in the allocation of resources among countries. The increased use of financing may imply an increase in the rate of inflation. An optimal international system should balance these various considerations. For discussion of efforts to design such a system, see Solomon (1982) and the documents of the IMF's Committee of Twenty (IMF 1974).

In a purely laissez-faire system, market borrowing instead of official reserves would be the source of financing to postpone adjustment. Fluctuations in market interest rates would determine the terms of trade between adjustment and financing. As is usual in market solutions, the wealthy are in a better position to negotiate terms on loans. By contrast, a more institutionalized system provides access to financing at lower rates to those with a weaker market position, with more conditions on the use of the funds. Evaluating the difference between two such systems is a complex task. For an attempt, see Jones (1983).

The Evolution of International Monetary Institutions

Between the close of the Napoleonic Wars and 1880, the international monetary system gradually

moved onto the gold standard, which was fully achieved during the period 1880–1914. Under the leadership of Great Britain, sterling operated as a vehicle currency during this period, allowing an efficient international payments mechanism to develop. The increasing substitution of bank deposits for currency allowed an ever-larger volume of payments to be supported by a gradually rising supply of gold. Despite the best efforts of the Bank of England and other central banks, periodic crises interfered with the continued convertibility of individual currencies. And the system was characterized by substantial fluctuations in employment and prices, albeit about a rising trend of employment with no trend in prices.

Following the First World War, gold convertibility was resumed on a limited basis, until the Great Depression of 1929–33 brought it to an end. A period of fluctuating exchange rates, competitive devaluations, and increasing use of trade restrictions to promote domestic employment ensued. It is generally believed that the economic difficulties of the interwar period were major factors bringing on the Second World War.

The Bretton Woods System

The United States and Great Britain took the lead in constructing the post-war international monetary institutions, with Harry Dexter White and John Maynard Keynes drawing up rival designs for the new system agreed at the Bretton Woods Conference in 1944. The Articles of Agreement of the International Monetary Fund provided for a system based on pegged, but adjustable, exchange rates and an institution which would lend reserve assets to countries that were having temporary difficulties in maintaining convertibility. Resort to floating exchange rates, competitive devaluations, and trade restrictions to promote domestic employment were explicitly to be avoided, in the light of the problems of the 1930s. Convertibility for current account transactions was promoted, while capital account convertibility was required only for those transactions necessary for financing current payments.

The lending power of the IMF was based on *quotas* of gold and domestic currency contributed by each member country. Only the gold was to be paid in initially, but, if the Fund needed convertible currency to lend out, it would obtain it from any member whose currency was considered strong enough to be *usable.* Members could borrow automatically up to the amount of the gold portion or *tranche* of the quota, but only on demonstration of balance of payments need, and thereafter they could borrow more subject to meeting conditions on economic and financial policies. For further discussion of IMF policies, see Williamson (1983), Kenen (2001), and Truman (2006).

The initial post-war problem involved the establishment of a payments system that would promote economic recovery and the growth of trade among the former combatants. The International Monetary Fund limited itself to establishing a set of agreed par values for pegged exchange rates which could promote the growth of trade, leaving the provision of loans and grants for economic recovery to the United States, the strongest economy. Under this system, which was a form of gold exchange standard, countries declared their par values in terms of the US dollar, which in turn was convertible into gold at $35 an ounce. Thus the dollar became the key currency of the system, and most foreign exchange reserves came to be held in the form of dollars. Within Europe, convertibility remained limited until 1958, and the European Payments Union was established to facilitate intra-European payments. The re-establishment of convertibility led to fears that the IMF might have inadequate resources to deal with the problems of large member countries. In 1962 the General Arrangements to Borrow were created, to enable the Fund to mobilize additional resources from its largest members, the Group of Ten.

With the recovery of the European economies in the 1950s and the achievement of convertibility in 1958, the US dollar became gradually overvalued relative to gold and other currencies. As Robert Triffin (1960) pointed out, the key currency system required the United States to continue to run balance of payments deficits in order to supply other countries with increased foreign exchange reserves. As it did so, the gold reserve of the United States became increasingly inadequate

to guarantee gold convertibility of growing US official dollar liabilities at $35 an ounce.

A variety of solutions to this problem were proposed, including the creation of an artificial reserve asset to substitute for dollars, an increase in the dollar price of gold, and the adoption of floating exchange rates. In 1968 the First Amendment to the Articles of Agreement of the International Monetary Fund permitted the creation of Special Drawing Rights (SDRs), which have twice been allocated to member countries in proportion to their existing quotas in the Fund. SDRs, when utilized, permit the user to acquire convertible currencies from other members, upon the payment of interest. They represent a centralized mechanism for increasing the stock of reserves. By the early 1970s the gold convertibility of the dollar was under increasing pressure, for a variety of reasons. In August 1971 the dollar was unilaterally set loose from gold. The Smithsonian Agreement of December 1971 attempted to save the Bretton Woods system by multilateral realignment of exchange rates, including a devaluation of the dollar against gold and a widening of the narrow bands of fluctuation permitted around the newly fixed values. Some members of the European Communities (EC) agreed to maintain narrower margins of fluctuations versus each other's currency, in an arrangement that became known as the 'EC Snake'. Despite these efforts, the revised Bretton Woods system lasted only a little more than a year.

Floating Exchange Rates

In March 1973, exchange rates of most of the major industrial countries began floating. At the same time, most developing countries continued to peg their currencies to the dollar or another developed country currency, and the EC maintained the 'Snake'. About this time, a major effort to reconstruct international monetary institutions on the basis of pegged exchange rates began under the auspices of the IMF's Committee of Twenty. This effort collapsed in 1974, in part under the impact of the quadrupling of world oil prices by the Organization of Petroleum Exporting Countries.

In Jamaica in January 1976, the Interim Committee of the Board of Governors of the International Monetary Fund agreed on a Second Amendment to the Fund's Articles of Agreement, ratifying the system of floating exchange rates. First, stability of exchange rates was to be sought through stability of underlying monetary and fiscal policies rather than through pegging. Second, floating rates should be subject to a process of 'firm surveillance' by the IMF. Third, it was hoped that the SDR would 'become the principal reserve asset', with the role of gold and the dollar being reduced. Fourth, the fixed official price of gold was abolished and one-third of the IMF's gold was disposed of. Acceptance of the status quo was all that could be accomplished. The result, according to Corden (1983), was an international laissez-faire system.

In 2005 some 88 countries made use of floating exchange rates, while 51 had pegged exchange rates of one type or another and 48 operated within currency unions with other countries.

Increased Capital Mobility, the Asian Crisis and Reform Proposals

Beginning in the 1970s, international capital mobility increased significantly, as middle-income developing countries found new access to foreign borrowing and industrialized countries increasingly opened production facilities in each others' markets. In the early 1990s, the IMF began discussions of a possible amendment that would promote capital account convertibility as an additional goal of the international monetary system, on the argument that improved allocation of capital would lead to increased economic growth. But a series of crises in emerging market economies interfered with this project, most notably the Asian financial crisis of 1997, followed by the Russian crisis of 1998 and the Argentine crisis of 2001. Each of these events was preceded by substantial capital inflows seeking higher returns, which overwhelmed underregulated and underprepared domestic economies and financial systems. The convertibility of affected currencies was often temporarily impaired (Black et al. 2006). In some cases the IMF was seen as creating a permissive environment prior to the crisis, followed by harsh demands for domestic reforms subsequently, in

attempts to restore confidence and bring an end to capital outflow.

A substantial body of criticism on one side argued that, by its willingness to provide large amounts of financing to countries in crisis, the Fund had created 'moral hazard', encouragement to over-borrowing and over-lending in expectation of a bailout (International Financial Institution Advisory Commission 2000). On the other side, others claim that the Fund by its harsh requirements for reform was stifling economic recovery and growth (Stiglitz 2002). Both of these viewpoints may have had some validity, but in a sense they cancel each other out (see Kenen 2001). The Fund itself proposed creation of an international Sovereign Debt Restructuring Mechanism to assist defaulting countries in negotiations with creditors (Krueger 2003). This was rejected in favour of a more modest approach encouraging the use of collective action clauses in bond indentures requiring minority bondholders to accept terms of repayment agreed to by a majority.

Another criticism of the IMF is that its voting shares and representation appear outdated, as compared with the changing economic importance of different groups of countries (Truman 2006). In particular, large emerging market economies such as China, India, and Brazil are underrepresented, while the European Union countries with 32 per cent of the voting power are overrepresented. Obviously, changes in representation are extremely difficult to achieve, but will still be necessary to remedy a situation in which the rich creditor countries that do not utilize the Fund's resources have disproportionate voting power relative to the debtor nations that have greater need for use of its facilities.

The 'New' Bretton Woods and Asian Monetary Cooperation

Following recovery from the Asian crisis of 1997, countries such as Korea, China, Malaysia, Taiwan and India sharply increased their accumulations of international reserves, as developing Asian countries in total raised their reserves (minus gold) from SDR 414 billion to SDR 1,039 billion between the ends of 1998 and 2004. China, Hong Kong and Malaysia in particular sought to maintain exchange rates pegged to the US dollar, while the other countries managed their floating exchange rates so as to avoid undue appreciation against the US dollar, accumulating enormous reserves in the process. An influential paper by Dooley et al. (2004) argued that this relationship was a new version of the old Bretton Woods system, whereby other countries pegged their exchange rates to the US dollar, enabling the United States to run large current account deficits, while the creditor nations increased their exports to the United States. Alternatively, the vastly increased reserve holdings of Asian countries could be regarded as a precautionary response to insure the availability of financing to avoid the prospect of another sharp adjustment, following the unpleasant experiences of the 1997 Asian crisis.

The combination of increased regional reserve holdings and recent bad experience with internationally supervised adjustment has led Asian countries to embark on steps towards regional monetary cooperation, culminating in the so-called Chiang Mai Initiative for regional currency swaps among the Association of South East Asian Nations (ASEAN) plus China, Japan, and Korea (see Park and Wang 2005). ASEAN members realized that the industrial countries of the Group of Ten had previously used currency swaps among central banks to lend each other money in times of crisis and thus avoid the need for borrowing from the IMF with its conditionality. With growing availability of reserves in Asia, the ASEAN + 3 concluded that they might similarly help each other out in future. Under the leadership of the Asian Development Bank, further steps are contemplated, possibly including an Asian Monetary Fund and an Asian Currency Unit.

The European Monetary Union

The enlargement and strengthening of the EC 'Snake' in 1978, which was in the process renamed the European Monetary System (EMS), gradually led to the creation of the European Monetary Union with a unit of account, the European Currency Unit (ECU). The objectives of the enlarged EMS were to reduce intra-European exchange rate fluctuations, to promote convergence of macroeconomic policies within

Europe, and to reduce European dependence on US monetary policies. Over a period of 15 years, the EMS succeeded in these objectives, at the cost of a series of exchange rate realignment crises culminating in a major collapse of the system in 1992–3, when the narrow margins (plus or minus $2\frac{1}{4}$ per cent) were expanded (to plus or minus 15 per cent). The crisis was brought on by a combination of increasingly rigid exchange rates within the system, increased capital mobility as a component of the Single Market programme of the European Union, and stresses brought on by the unification of East and West Germany.

In response to these factors, and to further strengthen the integration of European markets and achieve a more symmetrical sharing of decision making in monetary policy, the Maastricht Treaty ratified in 1993 brought into being in 1999 the European Monetary Union, with a single currency, the *euro*, with monetary policy controlled by a European Central Bank (ECB) in Frankfurt, Germany, replacing the currencies of the 12 member countries of the eurozone. While the euro has been quickly accepted as an international currency, in both the member countries and their neighbours, the relatively conservative operations of the ECB together with the constraints on member countries' fiscal policy embodied in the Stability and Growth Pact have proven controversial in the light of slow economic growth in the eurozone.

The euro is gradually becoming more important in international transactions and in the foreign exchange market as a rival to the US dollar. In 2006 the IMF redefined the SDR currency basket reflecting the importance of currencies in international trade and finance to be composed of 44 per cent US dollars, 34 per cent euro, 11 per cent Japanese yen and 11 per cent pound sterling, as compared with the previous weights of 45 per cent US dollars, 29 per cent euro, 15 per cent yen and 11 per cent pound sterling.

See Also

- ► Capital Controls
- ► Gold Standard
- ► International Capital Flows
- ► International Reserves

Bibliography

Bhagwati, J. 1978. *Anatomy and consequences of exchange control regimes*. Cambridge, MA: Ballinger.

Black, S. 1985. International money and international monetary arrangements. In *Handbook of international economics*, ed. R. Jones and P. Kenen, vol. 2. Amsterdam: North-Holland.

Black, S., C. Christofides, and A. Mourmouras. 2006. Convertibility risk: The precautionary demand for foreign currency in a crisis. *Annals of Finance* 2: 141–165.

Clark, P. 1970. Optimum international reserves and the speed of adjustment. *Journal of Political Economy* 75: 356–376.

Corden, W. 1983. The logic of the international monetary non-system. In *Reflections on a troubled world economy*, ed. F. Machlup et al. London: Macmillan.

Dooley, M., D. Folkerts-Landau, and P. Garber. 2004. The revived Bretton Woods system. *International Journal of Finance and Economics* 9: 307–313.

Eichengreen, B., and R. Hausmann, eds. 2005. *Other people's money: Debt denomination and financial instability in emerging market economies*. Chicago/London: University of Chicago Press.

Heller, H. 1966. Optimal international reserves. *Economic Journal* 74: 333–352.

IMF (International Monetary Fund). 1974. *International monetary reform: Documents of the committee of twenty.* Washington, DC: IMF.

International Financial Institution Advisory Commission. 2000. *Report*. United States Congress: Washington.

Jones, M. 1983. International liquidity: A welfare analysis. *Quarterly Journal of Economics* 98: 1–23.

Kelly, M. 1970. The demand for international reserves. *American Economic Review* 60: 655–667.

Kenen, P. 2001. *The international financial architecture: What's new? What's missing?* Washington, DC: Institute for International Economics.

Krueger, A. 2003. Sovereign debt restructuring: Messy or messier? *American Economic Review* 93 (2): 70–74.

Park, Y., and Y. Wang. 2005. The Chiang Mai Initiative and beyond. *World Economy* 28: 91–101.

Solomon, R. 1982. *The international monetary system, 1945–1981*. New York: Harper & Row.

Solomon, R. 1999. *Money on the move: The revolution in international finance since 1980*. Princeton: Princeton University Press.

Stiglitz, J. 2002. *Globalization and its discontents*. New York/London: Norton.

Triffin, R. 1960. *Gold and the dollar crisis*. New Haven: Yale University Press.

Truman, E. 2006. *A strategy for IMF reform*. Washington, DC: Institute for International Economics.

Williamson, J., ed. 1983. *IMF conditionality.* Washington, DC: Institute for International Economics.

I

International Monetary Policy

Paul De Grauwe

One of the main characteristics of the international monetary system is the absence of an international monetary authority (central bank) with policy making powers comparable to those central banks have at the national level. Whereas national central banks typically regulate domestic money markets in one way or another, there is no comparable authority to regulate international money markets. As a result, international monetary conditions will be the outcome of a decentralized decision-making process, in which market forces play a role together with the policies of a few important countries. Ultimately, therefore, international monetary relations will be influenced by the nature of the cooperation (or the lack of cooperation) among the central banks of the major countries.

There exist, of course, a number of international monetary institutions with important responsibilities. The most noteworthy are the Bank for International Settlements (BIS), and the International Monetary Fund (IMF). The latter has a major responsibility in providing credit to the countries with balance of payments and foreign debt problems. In addition, this role of the IMF has tended to increase during the last decade. Nevertheless it is fair to say that these institutions are far from having the powers and the responsibility a typical central bank has at the national level. It is also very unlikely that any of these institutions can be promoted to the position of a true world central bank in the foreseeable future.

The absence of a world central bank implies that the international monetary situation will be heavily influenced by the actions of the monetary authorities of the major countries. In fact, the nature of the domestic monetary policy regime in these countries is of crucial importance in the determination of the nature of the international monetary system. It is, therefore, useful to see how these domestic monetary policy regimes have changed over time, in particular in the United States which is the single most important country.

During the early postwar period the prevailing view in the US (and in other industrialized countries) was that the major responsibility of the central bank consisted in maintaining domestic price stability. This view, which originated in the writings of the classical economists, provided implicitly or explicitly the framework for monetary policy making in the major industrialized countries. The spillover of this view and of this policy attitude in the international sphere was a system of fixed exchange rates. The largest country, the US, was successful in stabilizing its price level. The other countries pegged their exchange rates to the US dollar, thereby also obtaining domestic price stability. As a result, during that period (which lasted roughly until the mid-sixties, and which is usually called the Bretton Woods System) the world experienced stable exchange rates and the absence of inflation. This arrangement, however, could only work satisfactorily if the countries pegging to the dollar were willing to subordinate their domestic monetary policies to the maintenance of a fixed dollar rate of their currencies. As the US policy was predicated on maintaining price stability, the willingness of the other countries to impose on themselves the discipline of a fixed exchange rate was great.

This situation began to change when views about the responsibilities of the central bank altered. Instead of being the guardian of a stable purchasing power value of money, the central bank was increasingly seen as an institution responsible for the stabilization of economic activity. This led to problems with the stabilization of the price level, and undercut the basis of the fixed exchange rate system. Inevitably, as countries used monetary policies to stabilize output, inflation rates became more variable and also more different across countries. In the end, fixed exchange rates had to be abandoned, and from the early Seventies on the industrialized countries allowed their exchange rates to float.

The responsibilities of the US in the fundamental change of the international monetary environment have been widely discussed (see e.g. Triffin

1968; Niehans 1974). By abandoning price stability as the major monetary policy objective, the US also transmitted inflationary shocks to those countries pegged to the dollar. As a result, these countries lost their willingness to maintain a fixed exchange rate with the dollar.

These changes in the domestic monetary policy regimes in the major industrialized countries have made exchange rates inevitably more variable than in the Bretton Woods period. This has also led to fundamental changes in the international monetary policy environment. Changes at three different levels can be identified.

A first change concerns the nature of the cooperation of the central banks. During the Bretton Woods period, the discipline of fixed exchange rates forced countries to coordinate their monetary policies closely. In fact, this coordination more or less automatically followed from these pegging arrangements. In addition, since pegging of the exchange rates occurred vis-à-vis the dollar, US monetary policies determined monetary conditions in the rest of the world. Thus, it can be said that the Bretton Woods system was a cooperative international monetary arrangement based on the leadership of the US.

The shift towards flexible exchange rates changed the nature of international monetary cooperation. First of all, as the movements of exchange rates tended to absorb monetary disturbances, the need to coordinate national monetary policies was generally felt to be less urgent. Second, when cooperation took place it tended to be of an ad hoc nature instead of automatic as in the Bretton Woods period.

It has been argued that this lack of explicit cooperative arrangements among the monetary authorities of the major industrialized countries is in itself a factor explaining the high volatility of the exchange rates observed since 1973 (see for example Williamson 1984). Certainly, this volatility of the exchange rates came as a surprise to most academic economists, who had been influenced by the conventional wisdom of the Sixties stressing that a flexible exchange rate system would make a smooth adjustment of external equilibrium possible (see Friedman 1953; Johnson 1967; Sohmen 1961). Those who read the old writers on the subject, however, knew better (see for example Bernholz 1982, for a survey of older views about flexible exchange rates.)

A second major change since the inception of flexible exchange rates concerns the nature of monetary interdependence between nations. Academic opinion prior to the Seventies was that a system of flexible exchange rates would make individual countries more independent in setting domestic monetary policies than a system of fixed exchange rates. In particular, it was thought that flexible exchange rates would allow countries to determine their own inflation rates, so that even when other countries followed inflationary policies, those countries that wanted it could insulate themselves from these foreign inflationary shocks.

These predictions of the merits of flexible exchange rates have been fulfilled only partially. It turned out to be true that countries can select their long-run inflation rates more or less independently if they allow their exchange rate to vary. Thus it was possible for countries such as Switzerland to have an inflation rate of only four per cent per annum during 1973–84 whereas the average in the industrialized countries was more than eight per cent per annum. On the other hand, it also appears that the short-term movements of inflation rates have been more correlated across countries during the floating rate period than during the period prior to 1973. Thus, although countries now have a higher degree of independence in selecting their long-run inflation rates, the yearly movements in these inflation rates turn out to be more dependent on outside price shocks than during the Bretton Woods period.

The reasons for this unexpected phenomenon are twofold. First, the occurrence of flexible exchange rates coincided with major supply shocks during the Seventies which tended to raise the rate of inflation in all countries. Second, the exchange rate regime experienced since 1973 was not a pure floating exchange rate system. Central banks continued to intervene heavily in the foreign exchange markets. They did this in (usually unsuccessful) attempts to mitigate the movements of the dollar. Thus, during the period 1973–78 the dollar declined substantially against

the other major currencies, and central banks bought massive amounts of dollars in order to stem its slide. This had the effect of expanding the money stocks in all these countries and tended to accelerate inflation. Exactly the opposite occurred during the period 1979–84 when the dollar experienced an unprecedented surge, and when central banks sold dollars and deflated their own money stocks.

This system of managed floating produced the curious result that monetary expansions and contractions which originated in the US were transmitted to the other industrialized countries as they would have been under a fixed exchange rate arrangement. And yet the dollar continued to be highly volatile, as these interventions in the dollar exchange markets failed to have much effect on the movements of the dollar. In a sense it can be said that the international monetary arrangement of the Seventies and the early Eighties combined the disadvantages of fixed and flexible exchange rates.

The shift to more flexible exchange rates produced a third major change in the international monetary policy environment. During the Bretton Woods period a major concern of monetary policy makers was control of the creation of international reserves. It was then widely felt that the mechanism of international reserve creation which was implicit in the gold-exchange standard was deficient. The rate of growth of the stock of international reserves (gold and dollars) did not correspond to the needs of an expanding world trade. In addition, the system had the disadvantage of leading to a disproportionate growth of the dollar stock relative to the stock of gold. As a result, a confidence problem arose concerning the ability of the US to maintain the gold convertibility of the dollar (a problem analysed by Triffin 1960, and Rueff 1961).

The 'liquidity problem' of the Bretton Woods system led to numerous schemes to manage the creation of international reserves (for a survey, see Grubel 1969, chapters 7 and 8). Without exaggeration it can be said that in those days the single most pressing issue of international monetary policy was thought to be this liquidity problem. Ultimately, this concern led to the creation of Special Drawing Rights, which were intended to substitute for gold, and which would enable the international community to regulate the creation of international reserves in a more rational way.

With the breakdown of the Bretton Woods system and of the gold exchange standard, these concerns about the creation of international reserves tended to fade away. Whereas in the Bretton Woods era the consensus was that the most important international monetary problem was how to create international reserves, there is now a growing consensus that the single most important issue faced by the international monetary system today is whether the degree of exchange rate variations has not become excessive. Concern that this may be the case has led some economists to propose schemes of coordination of monetary policies of the major industrialized countries, so as to stabilize exchange rates.

The need to come to such explicit cooperative agreements between central banks remains a controversial issue. There are essentially two schools of thought. The proponents of international monetary cooperation (e.g. McKinnon 1984; Williamson 1983) argue that the present flexible exchange rate regime leads to excessive movements of exchange rates, thereby making domestic macroeconomic management, and in particular the stabilization of the domestic price level, more difficult. In this view, cooperative agreements aimed at stabilizing the exchange rates must be given priority in order for countries to stabilize their economies in a more effective way.

A second school of thought turns the argument around and claims that domestic monetary stability comes first. In order to stabilize the exchange rate the monetary authorities of the major industrialized countries must follow more stable and predictable monetary policies. If this is done, the domestic price level can be stabilized so that the exchange rates can follow a more stable and predictable path. (Representative proponents of this view are Willett 1983; Haberler 1977.)

This debate has gone through many cycles in history. During the Twenties, after a period of strong fluctuations of the exchange rates, there was a widely held conviction that the paramount

task in the field of international monetary cooperation was stabilizing the exchange rates of the major currencies. This was seen as a first step towards the successful stabilization of domestic economies (see Clarke 1967, for a history of central bank cooperation during 1924–31). The whole cooperative effort underlying the Bretton Woods system was inspired by the same idea. During the 1960s, as major countries relaxed the monetary discipline needed to sustain a fixed exchange rate system, the view that domestic stability was a precondition for exchange rate stability gained respectability. In the early Seventies this view had become predominant among academic economists. Now, after many years of volatile exchange rate behaviour, the old view stressing the need to stabilize the exchange rates as a first step toward achieving domestic stability has regained some respectability.

The conflict between these two views, however, has not yet been settled. As a result, there is as yet no general agreement on how monetary policy should be conducted at the international level.

See Also

- ▶ International Finance
- ▶ Monetary Policy
- ▶ Supply Shocks in Macroeconomics
- ▶ Transfer Problem

Bibliography

Bernholz, P. 1982. *Flexible exchange rates in historical perspective*, Princeton Studies in International Finance No. 49. Princeton: Princeton University Press.

Clarke, S.V.O. 1967. *Central Bank Cooperation 1924–1931*. New York: Federal Reserve Bank of New York.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.

Grubel, H. 1969. *The international monetary system*. Harmondsworth: Penguin Books.

Haberler, G. 1977. The international monetary system after Jamaica and Manila. *Weltwirtschaftliches Archiv* 113: 1.

Johnson, H. 1967. Theoretical problems of the international economy. *Pakistan Development Review.* Reprinted in R. Cooper, *International finance.* Harmondsworth: Penguin Books, 1970.

McKinnon, R. 1984. *An international standard for monetary stabilization*, Policy Analyses in International Economics 8. Washington, DC: Institute for International Economics.

Niehans, J. 1974. Reserve composition as a source of independence for national monetary policies. In *National monetary policies and the international financial system*, ed. R.Z. Aliber. Chicago: Chicago University Press.

Rueff, J. 1961. Gold exchange standard a danger to the west. *The Times,* London, 27–29.

Sohmen, E. 1961. *Flexible exchange rates, theory and controversy.* Chicago: University of Chicago Press.

Triffin, R. 1960. *Gold and the dollar crisis*. New Haven: Yale University Press.

Triffin, R. 1968. *Our international monetary system*. New York: Random House.

Willett, T. 1983. Functioning of the current international financial system: Strengths, weaknesses and criteria for evaluation. In *International money and credit: The policy roles*, ed. G. von Furstenberg. Washington, DC: International Monetary Fund.

Williamson, J. 1983. *The exchange rate system*, Policy Analyses in International Economics, No. 5. Washington, DC: Institute of International Economics.

International Outsourcing

Deborah L. Swenson

Abstract

International outsourcing involves the import of intermediate inputs or services from unaffiliated foreign suppliers. While it implies that the production of a final product involves production activities in more than one country, this trade in intermediate inputs can be explained by traditional theories of international trade where countries have comparative advantage in different stages of production. However, since outsourcing relationships involve interaction with foreign partners, the choice of organizational form for these transactions is also influenced by industrial organization factors, such as search costs or contract incompleteness. This article discusses these

issues and the effects of outsourcing on the international economy.

Keywords

Comparative advantage; Factor price equalization; Foreign direct investment; Hold-up problem; Incomplete contracts; Intermediate inputs; International outsourcing; Matching; North–South economic relations; Offshoring; Search costs; Sunk costs; Trade costs; Thick markets; Transport costs; Vertical integration

JEL Classifications

F1

In complex production processes firms face a classic make-or-buy question: should they purchase parts, assembly or services from an outside vendor, or perform those tasks themselves? In the domestic context, the benefits and costs of vertical integration are already well understood. However, declines in international transport costs, advances in remote management technologies and improved communications technologies have brought an international dimension to this question, as they have enabled an increasing number of firms to engage in international outsourcing, purchasing parts, assembly or services from unaffiliated international suppliers.

As with trade in final products, international trade in intermediate inputs is shaped by international differences in comparative advantage that reflect cross-country differences in factor costs, or relative productivities for different stages of production. Hence, when firms decide where to complete the production activities required for the creation of a final product – design, materials extraction, parts production, and assembly – comparative advantage influences the ideal country placement for each production stage. However, if there are international frictions, such as international transport costs or tariffs, outsourcing imports will emerge only when the international outsourcing benefits stemming from comparative advantage exceed the costs associated with these international frictions.

Broad examination of international trade data, such as that of Hummels et al. (2001), or Yeats (2001) indicates that international trade in intermediate inputs has grown even more rapidly than the generally large growth in international trade since 1960. This trade in intermediate inputs represents both outsourcing purchases from unrelated suppliers and 'offshoring', which is the import of parts or services from related overseas suppliers, such as foreign subsidiaries. Yi (2003) argues that two key economic factors explain why recent declines in international trading costs have generated such an exceptional increase in intermediates trade. First, trade in intermediates increases along an extensive margin as declines in trade costs enable products that were previously produced domestically to be more profitably produced through an internationally integrated production process. Second, the effects of declining frictional costs are magnified when intermediates trade involves multiple border crossings, since the benefits of falling tariff or transportation charges apply to each border crossing involved in the creation of the final product.

Nonetheless, while international differences in international factor costs provide an incentive for international outsourcing, differences in international factor costs are not sufficient in themselves to guarantee that outsourcing relationships will develop. Thus, there are two strands in the literature on international outsourcing that explain firm choices. The first emphasizes limits on outsourcing that relate to search costs and matching, while the second focuses on the firm's choice of organizational form when contracts are incomplete.

When the search for an appropriate outsourcing partner is costly, firms will search for a foreign partner if the expected increase in profit generated by the search exceeds the sunk costs of searching for an international partner. Thus, Grossman and Helpman (2005) demonstrate that, when the cost of foreign search is particularly high, firms may choose domestic outsourcing in the high-wage home country over foreign outsourcing in the lower-cost foreign country. In addition, if the appearance of potential outsourcing partners is endogenous, the increased demand for partners in a particular location generates a market thickness externality. Since the

entry of potential partners increases the likelihood that searches will be successful, the increase in expected profits in thick markets increases the equilibrium number of searches in the market that becomes more densely populated with suppliers. As a result, search cost frictions, and the market thickness externalities they generate support an international equilibrium in which firms may be indifferent between searching for partners at home and searching for them abroad, even though wages and factor costs are not equalized across countries. In support of these ideas, Swenson (2005) finds that, while costs matter for some US outsourcing decisions, the cost sensitivity is largest for industries that are less capital intense and for industries that have thicker international markets for suppliers. The quality of country institutions, such as the strength and efficacy of a country's legal system, also influence the strength of market thickness externalities, since favourable country institutions increase demand for international outsourcing partnerships, thus increasing entry by outsourcing suppliers.

Even when comparative advantage is sufficiently strong to favour the overseas purchase of intermediate inputs, problems caused by contract incompleteness present a second reason why firms may not choose to engage in international outsourcing. In this case, firms may alternatively choose to integrate vertically with their foreign suppliers, or to set up foreign subsidiaries to conduct and support their purchase of overseas inputs and supplies. Firms are particularly likely to choose such 'offshoring' arrangements when problems arising from contract incompleteness are present in an industry that requires significant relationship-specific investments (Head et al. 2004; Qui and Spencer 2002). Antras (2003) argues that contract incompleteness will be more problematic in capital-intense industries: an idea that finds empirical support in his observation that the fraction of US imports that are intra-firm is higher for capital-intense industries and for trade with more capital- abundant countries.

Since the set-up of a foreign affiliate often involves substantial foreign direct investment expenditure, not all firms will choose offshoring over outsourcing for their international purchases of intermediate inputs. In this vein, Antras and Helpman (2004) show that, when firms are heterogeneous in their productivity, different sourcing strategies will coexist in equilibrium, as each firm chooses the sourcing method that maximizes its profits. Feenstra and Hanson (2005) provide further evidence of heterogeneous organizational choices in the case of Chinese processing trade, where organizational variation across Chinese industries and Chinese provinces supports their model of firm organization which is based on a property- rights description of the firm. Finally, thick market externalities may generate multiple outsourcing equilibria, as McLaren (2000) describes in a setting where independent input suppliers face a hold-up problem when they develop special components for a specific foreign purchaser of intermediate inputs. Here, an increase in market thickness, due to an increase in the number of final goods firms who search for suppliers, reduces the hold-up problem, since the bid of the next-closest purchaser increases.

Cross-country cost differences influence firms' international outsourcing decisions. In turn, the growth of international outsourcing may lead to changes in the international equilibrium. First, if country endowments differ dramatically, ordinary trade in final goods may narrow cross-country differences in factor rewards, but fail to bring about factor price equalization. Using traditional trade models, Deardorff (2001) shows that outsourcing may facilitate factor price equalization. Deardorff also shows that outsourcing may reduce a country's welfare if changes in international prices cause a terms of trade loss which reduces a country's gains from trade, as compared with the gains it reaped from trade in final goods only. However, outsourcing will not harm, and may even help, country welfare when international prices are unaffected.

The effect of outsourcing on international factor rewards depends crucially on the nature of the production process. In a three factor world where production of intermediate inputs involves the combination of capital with skilled and unskilled labour, Feenstra and Hanson (1996) demonstrate how outsourcing may exacerbate income inequality in all countries, where income inequality is

measured by the compensation of high-skilled relative to low-skilled workers. Intermediate goods are ordered according to their relative use of skilled to unskilled labour, while all intermediate inputs have equal capital cost shares. The production of the final good involves the costless assembly of the full range of intermediate inputs. In this framework, outsourcing brought about by capital flows from the Northern country to the Southern country reduces the relative cost of capital in the south, thus lowering the relative cost of producing each intermediate input in the south. This causes the skill intensity of southern production to rise as the country begins to produce an expanded range of intermediate inputs, which were previously completed in the North. From the South's perspective, the activities are more skilled-labour intense than their previous set of activities, while the activities were the least skilled-labour intense activities of those that the North produced. Thus, the shift in intermediates production increases the compensation of high-skill relative to low-skill workers in both locations since the relative demand for skilled workers rises in both the North and the South. At a firm level, Head and Ries (2002) observe that the skill level of Japanese workers in Japanese multinationals rose especially rapidly when the Japanese firms imported an increasing portion of their products from low-income, presumably labour-abundant countries. In the end, the fact that international outsourcing provokes such strong political concern reflects the fact that outsourcing, like international trade, has the potential to influence relative factor rewards.

See Also

- ▶ Foreign Direct Investment
- ▶ Incomplete Contracts
- ▶ International Trade Theory
- ▶ Vertical Integration

Bibliography

Antras, P. 2003. Firms, contracts and trade structure. *Quarterly Journal of Economics* 118: 1375–1418.

Antras, P., and E. Helpman. 2004. Global sourcing. *Journal of Political Economy* 112: 1375–1418.

Deardorff, A.V. 2001. Fragmentation in simple trade models. *North American Journal of Economics and Finance* 12: 121–137.

Feenstra, R.C., and G.H. Hanson. 1996. Globalization, outsourcing, and wage inequality. *American Economic Review* 86: 240–245.

Feenstra, R.C., and G.H. Hanson. 2005. Ownership and control in outsourcing to China: Estimating the property-rights theory of the firm. *Quarterly Journal of Economics* 120: 729–761.

Grossman, G.M., and E. Helpman. 2005. Outsourcing in a global economy. *Review of Economic Studies* 72: 135–159.

Head, K., and J. Ries. 2002. Offshore production and skill upgrading by Japanese manufacturing firms. *Journal of International Economics* 58: 81–105.

Head, K., J. Ries, and B.J. Spencer. 2004. Vertical networks and U.S. auto parts exports: Is Japan different? *Journal of Economics and Management Strategy* 13: 37–67.

Hummels, D., J. Ishii, and K.-M. Yi. 2001. The nature and growth of vertical specialization in world trade. *Journal of International Economics* 54: 75–96.

McLaren, J. 2000. 'Globalization' and vertical structure. *American Economic Review* 90: 1239–1254.

Qui, L.D., and B.J. Spencer. 2002. Keiretsu and relationship-specific investment: Implications for market-opening trade policy. *Journal of International Economics* 58: 49–79.

Swenson, D.L. 2005. Overseas assembly and country sourcing choices. *Journal of International Economics* 66: 107–130.

Yeats, A.J. 2001. Just how big is global production sharing? In *Fragmentation: New production patterns in the world Economy*, ed. S.W. Arndt and H. Kierzkowski. Oxford: Oxford University Press.

Yi, K.-M. 2003. Can vertical specialization explain the growth of world trade? *Journal of Political Economy* 111: 52–102.

International Policy Coordination

Paul R. Bergin

Abstract

Coordination among national governments as they formulate macroeconomic policies has been proposed as a response to global integration among national markets. Policy coordination may be beneficial by preventing the externalities created by policy spillovers, as

well as by promoting international risk sharing. The usefulness of coordination depends upon numerous characteristics of an economy, including the degree of openness in goods and asset markets.

Keywords

Asset market integration; Beggar-thy-neighbour; Bretton Woods system; Commitment; Coordinated solutions; European Central Bank; European Monetary Union; Exchange rate targets; Fiscal expansion; Globalization; Goods market integration; Information sharing; International capital flows; International migration; International policy coordination; Keynesianism; Labour market integration; Microfoundations; Monetary policy externalities; Nash solutions; Policy spillovers; Risk sharing; Sticky price; Terms of trade

JEL Classifications

F3; F23

Coordination among national governments as they formulate macroeconomic policies has been proposed as a response to global integration among national markets.

Awareness has grown over time of how national macroeconomies are interconnected in a global marketplace. Rising trade volumes indicate international integration among goods markets, large international financial flows indicate integration in asset markets, and highly visible immigration flows reflect increasing integration in national labour markets. Progressive globalization in the private economic sphere has prompted the question of whether public policy likewise should be global. Should the policies that nations use to manage their national macroeconomies be coordinated jointly with other nations? This is not a new question, and economists have voiced a variety of opinions and theories. Most academic economists have tended to be sceptical about the need for explicit international policy coordination.

To date there is limited coordination of macroeconomic policies in practice. Under the Bretton Woods arrangement of fixed exchange rates following the Second World War II until 1973, the monetary policies of member countries were constrained by the need to maintain an exchange rate target. If a national central bank were to attempt to increase the domestic money supply or lower domestic interest rates as a means of stimulating domestic production, this would tend to lower the value of its national currency relative to others and violate the fixed exchange rate agreement. Since the dissolution of this system in the 1970s, many nations learned to appreciate the resulting freedom to use their monetary policy to pursue domestic objectives.

Nonetheless, over the decades since the end of the Bretton Woods system, economic officials of major industrial countries periodically have met to discuss exchange rate intervention and options for monetary and fiscal policies. Examples include the Plaza and Louvre accords in the 1980s. Without binding public agreements, it is not clear how much coordination takes place at such meetings, and the function served by them may simply be sharing information regarding policy intentions. In some regions of the world, a subset of countries have taken steps on their own to more formally coordinate their policies. The most dramatic form of international macroeconomic policy coordination of late has been the formation of the European Monetary Union in 1999. Eleven initial member countries ceded sovereignty over national monetary policy to a European Central Bank, where a single monetary policy must be agreed upon for the whole region.

The opinions of academic economists on the advisability of policy coordination have varied over time, largely in response to the introduction of new tools of economic analysis. Milton Friedman (1953) and others recommended against explicit coordination, suggesting that private market forces could be trusted to achieve a desirable outcome. In particular, exchange rate movements could serve a useful function of insulating countries against the macroeconomic shocks of their neighbours. In contrast, economists of the 1970s and 1980s were able to find theoretical rationales for policy coordination, using Keynesian models that featured frictions that prevented economic markets from operating efficiently on their own.

Finally a renewed interest in the subject since 2000, employing models with more microeconomic foundations, has produced new theoretical reasons to question the usefulness of policy coordination.

The rest of this article considers two primary motivations for policy coordination: preventing policy spillovers and promoting pooling of international risk. The article discusses each motivation in turn along with its limitations.

Policy Spillovers

One motivation for policy coordination is the possibility that the effects of policy spill over national borders to affect the macroeconomies of trading partners. For example, suppose there is a global shock that lowers global demand below some desirable level, such as a wave of pessimistic expectations that lowers investment expenditure. This might be undesirable, to the degree the excess inventories may lead to recession, with a scaling back of production and lower levels of employment. Keynesian theory indicates that one way policymakers can combat such a shortfall in demand is through expansionary fiscal policy, with a rise in government expenditure or a cut in taxes to stimulate private consumption demand. However, globalization affects this policy prescription. National policymakers may fail to respond if they fear that some of the benefit will leak abroad: a fiscal expansion may lead to a currency appreciation, making domestic goods less competitive than foreign goods. As a result, some of the increase in demand generated by domestic government debt will be used to purchase foreign goods and employ foreign workers.

Coordination of policymakers across countries may provide a way of eliminating the problem created by this externality. If a mechanism of coordination existed to make sure that all countries symmetrically expanded government spending, each government could be reassured that it would benefit from spillovers of demand from abroad, to compensate for the negative spillover of demand leaking abroad. A coordinated global fiscal expansion, the theory says, is an effective way of combating a global shortfall in demand.

Externalities also apply to monetary policy. Monetary expansions tend to cause currency depreciations that make domestic goods more competitive compared to foreign goods. The use of such policy to shift demand from foreign goods toward home goods to raise domestic production at the expense of lower foreign production is labelled ‘beggar-thy-neighbour’. One might imagine repeated rounds of such policies, with each country progressively increasing money supply to regain competitiveness. In the end competing policies will have no net effect on the exchange rate and competitiveness, but the net rise in the money supply of each country would produce the undesirable outcome of excessive inflation. Coordination agreements may commit countries to avoid such policy outcomes; they may agree to forswear beggar-thy-neighbour policies if there is a credible commitment from other countries to do the same. The end result is a better outcome for all.

The spillover argument in favour of coordination clearly depends on the degree to which the private economies are interdependent internationally. Consider goods market integration. If exports tend to be a small fraction of a country’s GDP, a currency depreciation raising exports a certain percentage will have a small effect on GDP in absolute terms. The international implications of any policy just wouldn’t matter very much. Asset market integration also has been found to be important. If asset markets do not view government debt issued by different countries as equivalent, then a fiscal expansion that raises the issue of debt in one currency could cause a currency depreciation rather than an appreciation, reversing the direction of the fiscal spillovers described above.

Policy spillovers and strategic interactions of policymakers are topics introduced in research by Hamada (1974), Oudiz and Sachs (1984), and Canzoneri and Gray (1985). When a Keynesian theoretical model embodying the spillover arguments above was quantified by Oudiz and Sachs (1984), it was found that the gains from coordination were too small to justify the effort. US merchandise exports to Europe at the time amounted

only to 1.6 per cent of US GNP. As a result, the gains from coordination were estimated at only about 0.5 percentage points of GDP for the United States. The lesson was that since international integration was actually quite low, there was little or no role for policy coordination. A question that is addressed later in this article is whether this conclusion continues to hold in a progressively more globalized and integrated world.

Policymaker Objectives

The relevance of policy spillovers has been qualified by recent research that studies the objectives of policymakers; see Obstfeld and Rogoff (2002), Corsetti and Pesenti (2005), and Canzoneri et al. (2005). This research features microeconomic foundations to describe the behaviour of consumers, workers, and producers. One benefit of deriving consumer behaviour from assuming they are trying to maximize a particular utility function is that this utility function provides a natural metric by which to evaluate the benefits of alternative policies. Further, it facilitates predictions about how policymakers will act, on the assumption that their behaviour is driven by the goal of improving the welfare of private consumers. For example, one might assume that the policymakers of each country act independently to maximize the utility of citizens in their own country. This 'Nash' solution can be contrasted with a coordinated solution, where an international coordinator chooses the policies of all the countries jointly to maximize the sum of utility of citizens across countries. Only if the outcome of the latter coordinated solution supersedes that of the independent Nash solution is there a clear motivation for international policy coordination.

Consider a simplified theoretical world of two countries populated by representative agents that consume and produce. Production involves labour supplied by these agents, combined with technology that is subject to uncertain shocks each year. Suppose these economies exhibit a market imperfection in the form of prices that must be set ahead of time and that cannot change in response to surprise fluctuations in productivity. Given this environment, imagine there is a negative productivity shock that lowers the level of output. In contrast to the argument of the previous section, it no longer is clear that a policymaker should respond by trying to restore output to its previous level by stimulating demand. This would make the welfare of the citizens even worse, because it would force them to work harder during periods where their labours are less rewarded. Instead, utility is made highest by using monetary policy to replicate the outcome of an economy that is free from the sticky-price market imperfection. In this flexible-price version of the world, citizens would choose to work and consume less during periods of low productivity, and choose to work more and acquire wealth during periods when productivity shocks are favourable.

Although it may seem counter-intuitive, a policymaker wishing to maximize the welfare of his or her citizens in such an economy often will contract the money supply when output falls due to the productivity shock. This has the effect of raising the relative price of home goods and reducing demand and hence production. The outcome of this Nash game differs from the outcome described in the previous section, and does not involve any beggar-thy-neighbour strategy. The domestic policymaker is perfectly capable of replicating the flexible price outcome by the appropriate application of domestic policy. Under certain conditions to be discussed below it turns out that the coordinated solution is identical to that for the Nash solution above. If policy in the two countries were dictated by a central coordinator trying to eliminate all externalities, the set of policies he would prescribe for each country would be identical to the policies that each country would have chosen independently. In this world the spillover argument fails to apply, and there is no benefit from international policy coordination.

International Risk Sharing

A second type of motivation offered for international policy coordination is the possibility that countries can benefit by mutually insuring each other against the effects of shocks. Ideally private

asset markets would include trade in securities contingent on the incidence of shocks, which households could use to insure themselves. For example, in the case of a fall in productivity and output in just one country, such securities would require a transfer of wealth from abroad to this country as a way of buffering the level of consumption despite the fall in domestic production. International trade in equities could serve this function. Suppose the residents of two countries each own half of the firms in the other country's stock market, and they thereby have claim to half of the output of each other's total production. If a productivity shock lowers the output of the home country but leaves the foreign country unaffected, when each country sends half of its respective production to the other country, this implies a net payoff from the foreign country to the home country. This transfer effectively spreads the impact of the productivity shock over the consumption levels of both countries and acts as a type of insurance. However, in the absence of a private market for such securities, there may be a role for policy coordination to replicate these insurance benefits.

For example, consider again the story above of a negative shock to productivity in one country. Another motivation for the policymaker to employ a contractionary monetary policy is to raise the value of the domestic currency, in order to raise the relative price of its exports to imports, the terms of trade. By making home goods more valuable, he or she raises the revenue from export sales abroad, transferring wealth to the affected country. The ability to manipulate the exchange rate to transfer wealth from the foreign to home country clearly could present a temptation to pursue beggar-thy-neighbour policies. But in the hands of a central policy coordinator, this becomes a means of making transfers between countries when useful for insurance purposes.

Note that a coordinated policy motivated by the objective of risk sharing might in principle conflict with the motivation for coordination laid out in previous sections. There is no reason to suppose that the degree of monetary contraction needed to transfer enough wealth to pool risk is also that degree needed to discourage production to the level consistent with flexible prices. That is, it may not be possible to use policy coordination to offset two economic distortions at the same time, the sticky-price and imperfect risk-sharing distortions. This is a point emphasized in the influential work of Obstfeld and Rogoff (2002).

Extensions to more Realistic Economies

While recent research has noted additional theoretical rationales for coordination, this may not change the conclusion that the gains are too small quantitatively to justify the effort. When Obstfeld and Rogoff (2002) calibrate with reasonable parameter values a model that combines imperfect risk sharing with nominal rigidities, it does find there is some positive gain from a coordinator choosing a policy as opposed to each country optimizing separately. But the additional benefit from coordination is small. As long as policymakers act wisely to replicate flexible price outcomes in their domestic economy, the benefit of coordinating with foreign countries is smaller by an order of magnitude. Several features of the theoretical economic environment are key to this result. Clearly key is the supposition that policymakers will act in a manner to maximize the welfare of their residents when given the freedom to do so. But also essential are assumptions about the behaviour of consumers, such as the willingness to substitute across home and foreign goods to maintain their level of utility, and a desire to smooth consumption levels over time.

As progressively more realistic economic environments are explored, the list is augmented of economic features that affect the decision to coordinate. One such feature is the nature of price stickiness. When exporting firms set their prices, many will set them in the currency of the buyer's market. If prices are sticky in the local currency, any fluctuations in the nominal exchange rate will have no effect on the price that consumers face in the market. So any attempt to use monetary policy to manipulate the terms of trade as an insurance device will fail. As demonstrated in Devereux and Engel (2003), local currency pricing kills off a primary motivation for policy coordination as

well as the temptation to pursue beggar-thy-neighbour policies in a sticky-price world.

On the other hand, some other realistic economic features tend to augment the benefits of coordination. These include the reliance on imported goods as intermediates in the domestic production process, in which case random fluctuations in the exchange rate can severely disrupt domestic production. Such issues are likely to be most important for small economies, especially those that specialize in assembly operations of imported components for final export. Another relevant feature is the presence of nontraded goods. If the productivity shocks hitting the nontraded sector differ from the traded sector, it can become difficult for international trade in asset markets to insure against them. Calibrating and simulating models with these more realistic features indicates that it is possible for some economies to benefit substantially from policy coordination (see Tchakarov 2004).

In sum, the size of benefits from coordination depends on a number of key characteristics of economies. These include how developed asset markets are, how responsive trade flows are to relative prices, how important it is to households to smooth their consumption levels over time, how imports are used, and how sticky prices are set. Whether policy coordination is worthwhile for a country depends largely on the individual characteristics of that country.

Openness Reconsidered

While the discussion above has offered two motivations for policy coordination, namely, risk sharing and price stickiness, a revealing distinction between the two is how they are affected by openness and globalization. Consider first openness in the form of international economic integration in goods markets. Goods trade itself may have built-in mechanisms that can help insure a country against country-specific output shocks. For example, if a country is hit by a fall in its production, the relative scarcity of home goods would induce a rise in their relative price. Depending on consumer preferences, such as a type implying constant expenditure shares over home and foreign goods, this terms-of-trade effect will be able to compensate home agents for the fact they have a smaller quantity of home goods. In particular, they will be able to import more foreign goods in exchange for the smaller quantity of home exports, and thereby enjoy a comparable level of overall consumption and utility as the foreign country. This means that goods markets potentially can do the job of pooling risk internationally without the need for an international policy coordinator.

This conclusion stands in sharp contrast to earlier literature. Recall that Oudiz and Sachs (1984) concluded that the need for coordination was small precisely because the degree of goods trade was small. But here we conclude that the need for coordination is small when goods market integration is high.

Consider also the implications of integration in asset markets. In the limiting case where asset markets were complete, with assets to insure against all shocks, private agents would be able to pool the risk of asymmetric shocks internationally on their own. Again, if private markets pool risk, there is no need for policy coordination to serve this function. Clearly the world remains far from complete asset markets, but international trade in equities is definitely on the rise, and international capital flows of various types have ballooned. One gets the impression that international integration has progressed faster in asset markets than in goods markets, so that this type of integration may be more important.

Nevertheless, integration in both markets works in the same direction here. A high level of integration, be it in either asset markets or goods markets, indicates there is less need for explicit international policy coordination to pool national risks. Contrary to the predictions of some analysts, as the age of globalism progresses we might see less pressure for international policy coordination rather than more.

See Also

► International Finance
► International Monetary Institutions

▶ International Real Business Cycles
▶ Macroeconomic Effects of International Trade
▶ Monetary and Fiscal Policy Overview

Bibliography

Canzoneri, M.B., and J. Gray. 1985. Monetary policy games and the consequences of non-cooperative behavior. *International Economic Review* 26: 547–564.

Canzoneri, M.B., R.E. Cumby, and B.T. Diba. 2005. The need for international policy coordination: What's old, what's new, what's yet to come? *Journal of International Economics* 66: 363–384.

Corsetti, G., and P. Pesenti. 2005. International dimensions of optimal monetary policy. *Journal of Monetary Economics* 52: 281–305.

Devereux, M., and C. Engel. 2003. Monetary policy in the open economy revisited: Price setting and exchange rate flexibility. *Review of Economic Studies* 70: 765–783.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in positive economics*. Chicago: University of Chicago Press.

Hamada, K. 1974. Alternative exchange rate systems and the interdependence of monetary policies. In *National monetary policies and the international financial system*, ed. R. Aliber. Chicago: University of Chicago Press.

Obstfeld, M., and K. Rogoff. 2002. Global implications of self-oriented national monetary rules. *Quarterly Journal of Economics* 117: 503–535.

Oudiz, G., and J. Sachs. 1984. Macroeconomics policy coordination among the industrial countries. *Brookings Papers on Economic Activity* 1984(1): 1–64.

Tchakarov, I. 2004. *The gains from international monetary cooperation revisited*, IMF working paper no. WP/04/1. Washington, DC: International Monetary Fund.

International Real Business Cycles

Mario J. Crucini

Abstract

International business cycle research seeks to summarize the statistical properties of worldwide economic fluctuations and model them as the outcome of purposeful decisions by individuals, firms and policymakers who react to changes in their economic environment and an uncertain future. The focus is on identifying the sources of fluctuations and how interactions of economic actors play out in terms of cyclical movements in variables such as gross domestic product. The term 'real' indicates a sub-area of the research programme that focuses on non-monetary dimensions such as changes in productivity and fiscal policy rather than in the money supply and monetary policy.

Keywords

Adjustment costs; Armington aggregator; Asset market structure; Business cycle measurement; Business cycles; Capital accumulation; Capital utilization; Consumption function; Consumption smoothing; Current account; Euler equations; Expectations; Factor mobility; Imperfect competition; Incomplete markets; International finance; International real business cycles; International taxation; International trade; Intertemporal substitution effect; Law of large numbers; Law of one price; Nominal exchange rates; Open economy analysis; Permanent-income hypothesis; Purchasing power parity; Rational expectations; Real business cycles; Real exchange rates; Risk sharing; State space models; Sticky prices; Terms of trade; Total factor productivity; Trade costs; Wealth; World business cycle

JEL Classifications

F4

International Real Business Cycles

Business cycles are the recurrent fluctuations of national output relative to its long-term growth trend. The qualitative features of these fluctuations are common to virtually all economies, with their quantitative properties differing somewhat across countries and time periods. Modern research seeks to summarize the statistical properties of business cycles and formally model them as the outcome of purposeful decisions by individuals and firms who react to changes in

their economic environment and an uncertain future. Whereas closed-economy analysis focuses on responses to domestic shocks and policy actions, open economy analysis *adds* to this international policy interaction and spillovers of foreign shocks to the domestic economy. The term 'real' indicates a sub-area of the business cycle research programme that focuses on non-monetary dimensions such as changes in productivity, taxes and government spending, rather than changes in the money supply and monetary policy.

Measuring International Business Cycles

What may be surprising to the uninitiated is the controversy surrounding business cycle measurement itself. Measures most often cited in the press are the calendar dates of business cycle peaks and troughs. In the United States, these dates are identified by the Business Cycle Dating Committee at National Bureau of Economic Research. A committee affiliated with the Center for Economic Policy Studies serves the same function for Europe. The logic of the methods used by both committees dates back to the classic contribution of Burns and Mitchell (1946), pioneers of formal business cycle measurement.

In academic work, economists favour econometric methods in which the logarithm of real gross domestic product, y_t, is decomposed into a growth trend, $y_{g,t}$, and a business cycle component, $y_{c,t}$:

$$y_t = y_{g,t} + y_{c,t}. \tag{1}$$

A large applied econometrics literature achieves trend and cycle decompositions by applying identifying assumptions on the innovations to the trend and cycle components of aggregate output. See, for example, Beveridge and Nelson 1981; Cochrane 1994; Crucini and Shintani 2006; Stock and Watson 2005. Here we employ the Hodrick and Prescott (1997) filter to achieve this decomposition since it is widely used in the literature. The Hodrick–Prescott filter provides a smooth estimate of the growth trend, $y_{g,t}$, and the cycle is computed as the difference between the growth trend and the original series.

Figure 1 displays the business cycle component of the logarithm of gross domestic product for eight industrialized countries: Australia, Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States. As is evident, business expansions and contractions are persistent. One also sees common features such as the emergence of a recession in the 1980s simultaneously in most countries.

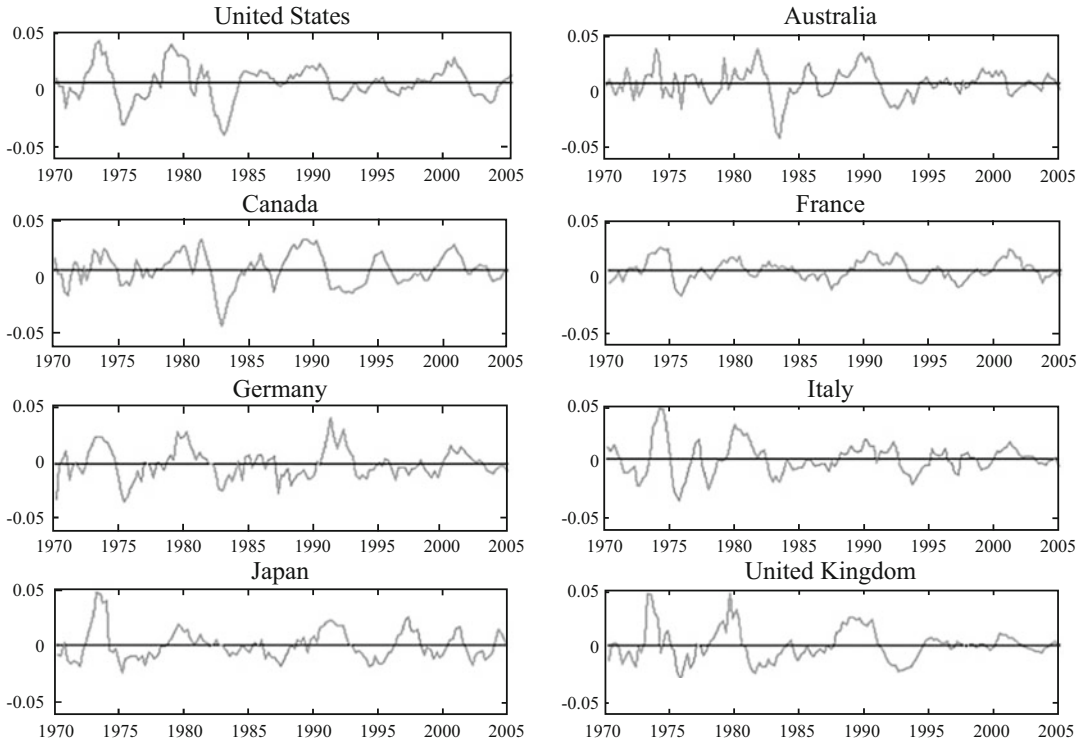
We organize our discussion of business cycle facts around two equations. The first is the national income and product accounts (NIPA) accounting identity. (The OECD data satisfy this identity when changes in inventories and a statistical discrepancy are included. We subtract these two items from output when we perform the variance decomposition of output from the expenditure side.)

$$Y_t = C_t + I_t + G_t + (X_t - M_t). \tag{2}$$

In words: the amount of output produced in the home country equals the sum of its uses in domestic private consumption and investment, C_t and I_t, government spending, G_t, and exports, X_t. Imports are deducted to avoid double counting since they are already counted in the other expenditure components.

The variables have been ordered in terms of the fraction of output accounted for by each component. Averaged across time periods and countries, consumption accounts for about 58 per cent of output and investment accounts for 23 per cent, while the percentages for government consumption, exports and imports are almost identical, at 18 per cent, 19 per cent, and 19 per cent, respectively. With the exception of exports and imports, the ratios differ modestly across industrialized countries when long-time averages are taken. We use Eq. (2) below to perform an *expenditure-side* decomposition of output variability.

The second relationship is a theoretical construct. The prototype model assumes that output is produced with two inputs, capital and labour. The production function relating inputs to outputs usually takes the form:



International Real Business Cycles, Fig. 1 Business cycle component of the logarithm of gross domestic product for eight industrialized countries, 1970–2005 (*Source*: OECD Quarterly National Accounts, CD-ROM and author's calculations)

$$Y_t = A_t K_t^{\alpha} N_t^{1-\alpha} \tag{3}$$

where A_t is total factor productivity, K_t is the stock of physical capital in place at time t, N_t is total hours of input at time t. The exponent $1 - \alpha$ measures the share of national income paid to labour (salaries and wages) since labour is paid its value marginal product in the model.

Taking logarithms of Eq. (3) provides the basis for the second variance decomposition:

$$y_t = a_t + k_t + (1 - \alpha)n_t. \tag{4}$$

We compute a_t as a residual, setting $\alpha = \frac{1}{3}$ (the share of capital income in national income) and using standard measures of physical capital and aggregate hours, as the inputs on the right-hand-side of the equation. We call this our *production-side* decomposition.

Table 1 contains business cycle statistics for each country using data from the first quarter of 1970 to the first quarter of 2005. Beginning with the variance of the cycle itself, we see that the United States has the most variable business cycle, with a standard deviation of 1.58 per cent per quarter, while France, at the other end of the scale, has a standard deviation of only 0.91 per cent. Australia, Canada, Germany, Italy, Japan and the UK have remarkably similar volatility, in the range of 1.32–1.48 per cent.

Turning to the details, we see that investment and trade flows are much more variable than output; consumption is less variable than output while government spending is the least variable. There are some quantitative differences across countries, but the rankings are robust.

The correlation of variables with output indicates the cyclicality of a variable. If the correlation is positive, the variable is said to be pro-cyclical: on average, it rises when the economy is in an expansionary phase and falls when the economy is in a contractionary phase. All

International Real Business Cycles, Table 1 Business cyclical properties of eight industrial countries, 1970Q1–2005Q1

	US	Australia	Canada	France	Germany	Italy	Japan	UK
Std. dev. of output	1.58	1.32	1.46	0.91	1.36	1.43	1.35	1.48
Panel A. Standard deviations relative to output								
Consumption	0.80	0.77	0.79	0.97	0.87	0.93	0.92	1.14
Investment	2.85	3.41	2.83	3.11	2.59	2.29	2.36	2.49
Government	0.54	1.26	0.78	0.78	0.86	0.55	0.92	0.72
Exports	2.68	3.00	2.66	3.11	3.00	2.71	3.21	1.97
Imports	3.26	4.83	3.16	3.95	2.32	3.24	4.31	2.54
Savings	4.46	4.88	3.72	4.07	3.70	3.03	2.52	4.17
Productivity	0.56	0.76	0.64	1.04	0.72	0.94	0.68	0.80
Capital	0.39	0.43	0.42	2.87	1.11	1.16	0.55	0.35
Labour	0.83	1.01	0.94	0.65	0.66	0.65	0.62	1.19
Panel B. Correlation with own-country output								
Consumption	0.85	0.42	0.82	0.71	0.67	0.75	0.79	0.79
Investment	0.95	0.78	0.60	0.83	0.81	0.79	0.93	0.66
Government	−0.18	0.07	− 0.15	− 0.20	0.10	−0.03	0.04	−0.19
Exports	0.42	0.11	0.67	0.72	0.62	0.24	0.05	0.48
Imports	0.81	0.45	0.73	0.80	0.74	0.70	0.62	0.68
Savings	0.88	0.86	0.89	0.81	0.82	0.82	0.84	0.74
Productivity	0.84	0.67	0.77	0.45	0.76	0.87	0.91	0.58
Capital	0.26	0.20	− 0.14	0.25	0.26	−0.08	0.37	0.09
Labour	0.90	0.68	0.84	0.67	0.81	0.47	0.75	0.67
Net export ratio	−0.44	− 0.32	− 0.09	− 0.28	0.08	−0.38	− 0.41	−0.30
Correlation of savings and investment	0.63	0.44	0.67	0.58	0.44	0.80	0.47	0.83

Notes: All variables except the net export ratio are the Hodrick–Prescott cycle components. All nominal variables are deflated by the Gross Domestic Product Deflator (*Source*: OECD Quarterly National Accounts, CD-ROM)

variables except government spending and the net export ratio are strongly pro-cyclical, consumption and investment particularly so. In a statistical sense, government spending seems to provide some stabilization by virtue of its low variability and near-zero correlation with the cycle. Imports are consistently more highly correlated with domestic output than are exports. This makes economic sense since import demand is influenced by domestic income while export demand depends on potentially diverse income developments across a country's trading partners.

On the production side of the equation, capital is less cyclically variable than either productivity or labour input (a notable exception is France). The ranking of the variability of labour input relative to productivity is ambiguous.

Variance Decompositions

The variance decomposition of output from the expenditure side or production side is computed as:

$$std(y) = \sum_z s_z \cdot std(z) \cdot corr(z, y) \qquad (5)$$

where s_z is either the expenditure share or the production share for variable z (productivity gets a weight of one), $std(z)$ is the standard deviation of component z over the cycle and $corr(z, y)$ is the correlation between component z and income. The variance decomposition is exact in levels, but approximate in logs, because the NIPA identity involves levels. The variance decomposition is exact on the production side because of the log-linearity of the production function.

On the expenditure side consumption and investment account for about 95 per cent of the cyclical variation in aggregate demand. There is no consistent ordering of their relative importance. The reason for consumption's impact is that about two-thirds of aggregate demand is accounted for by this component. While investment is a paltry 23 per cent of aggregate demand, it is about twice as variable as consumption and therefore exerts an influence on the cycle larger than its expenditure share would suggest. Imports are often as important as consumption or investment, while the contribution of exports is not robust across countries. However, since imports and exports enter the national income and product identity with opposite signs, they tend to cancel out. Fluctuations in government spending contribute little to the cycle, for three reasons. First, government spending accounts for a relatively small amount of aggregate demand, close to the investment and trade shares and much lower than that of private consumption. Second, government spending is typically less variable than output. Third, the correlation between government spending and output is close to zero, on average. (In periods of war, such as the Second World War, the picture is very different since government spending is a much larger fraction of output and is strongly pro-cyclical.)

To turn to the *production side*, total factor productivity and changes in labour input account for virtually all of the cyclical variation in output (the cross-country average contribution of these two combined is 95 per cent). This is because each of these variables is highly variable and highly correlated with output, much more so than is true of the physical capital stock. Moreover, capital's share in income is exactly one-half that of labour's, reducing its influence relative to labour. While productivity and labour have a comparable influence, the source of the influence differs. Labour input is more variable than productivity, but gets a weight of two-thirds, less than the unit coefficient on total factor productivity (see Eq. (4)).

It should be stressed that, while these accounting-based decompositions are useful in framing the discussion, they do not tell us what the underlying sources of business cycles are. To see this, consider the distinction between choice variables and exogenous variables. In the prototype real business cycle model, productivity is the only exogenous source of economic change, all other variables are responding optimally to this variable. The model, then, tells us that productivity variation accounts for all of business cycle variation and the various facets of how this plays out across macroeconomic aggregates reflect the choices made by individuals, firms and governments, in response to these productivity changes.

Thus, in practice, there is a subtle link between exogenous impulses and endogenous responses to them. For example, Imbs (1994) introduces variable capital utilization into the model described above. Since capital utilization is not part of what we are measuring in our physical capital stock series, we incorrectly allocate variation in capital utilization to productivity. It is natural to think that this leads us to overestimate the role of productivity. Baxter and Farr (2005) show, however, that, when one moves from a model with constant utilization of capital to one with variable utilization, the response of the economy to a productivity change of a fixed size is larger when utilization is variable than when it is fixed. This moves the bias in the other direction. The lesson here is that theory and measurement work best in concert to achieve the most accurate possible attribution of economic variance.

International Dimensions of the Business Cycle

We turn, now, to key international facets of business cycles: (*a*) the current account balance, (*b*) international business cycle co-movement and (*c*) relative price determination.

The Current Account

An important goal of international business cycle research is to improve our understanding of the time path of the current-account balance or the trade balance. International trade focuses on the direction and composition of trade and often assumes balanced trade. International finance

focuses on the current account, modelling the dynamics of savings and investment over time. Since the business cycle involves time variation, it is natural to emphasize the international finance perspective.

The current account equals the difference between savings and investment. National savings is the sum of private savings and public savings. Private savings is the difference between disposable income and private consumption while public savings is the difference between tax revenue and government expenditure.

$$\begin{aligned} CA_t &= S_t - I_t S_t \\ &= \underbrace{(Y_t - T_t - C_t)}_{\text{private saving}} + \underbrace{(T_t - G_t)}_{\text{public saving}}. \end{aligned} \tag{6}$$

In a closed economy, of course, the current account is identically equal to zero – each dollar of savings must be allocated to domestic investment. An open economy, freed from this constraint, rarely finds itself with a current account balance; when current savings fall short of (or exceed) current investment levels, a current account deficit (or surplus) obtains. Feldstein and Horioka (1980) vividly demonstrated that, when the data are averaged over long periods of time, savings and investment rates are highly positively correlated – countries with higher than average savings rates tend to have higher than average investment rates. Business cycle correlations of saving and investment tend to be lower than the Feldstein–Horioka values, suggesting that large deviations in the current account are transitory. The correlation of national saving and national investment over the cycle ranges from a high of 0.80 in Italy to a low of 0.44 in both Australia and Germany (see Table 1).

International Business Cycle Co-Movement

International co-movement may be expressed in different ways. Kose et al. 2003, among others, use state space models in which there are world, country and idiosyncratic factors in the income process as well as in each component of aggregate demand. This method avoids an arbitrary choice of numeraire and helps to identify what economists refer to as the 'world business cycle'. Here we use the correlation of a foreign variable with its US counterpart. As is evident in Table 2, positive movements of foreign variables with their US counterparts are the rule rather than the exception. In terms of rankings, output tends to be more correlated than the components of aggregate demand; investment and government spending have particularly low international correlations. The rankings are more ambiguous in a statistical sense and for a broader range of countries than Table 2 suggests; see Ambler, Cardia and Zimmerman (2004).

International Real Business Cycles, Table 2 International business cycle co-movement correlation with US counterpart, 1970Q1–2005Q1

	Australia	Canada	France	Germany	Italy	Japan	UK
Output	0.46	0.71	0.36	0.42	0.32	0.43	0.64
Panel A. Demand side							
Consumption	− 0.09	0.53	0.37	0.37	0.01	0.35	0.50
Investment	0.29	0.16	0.25	0.47	0.15	0.42	0.40
Government	0.22	0.29	− 0.04	0.05	− 0.01	0.07	0.06
Exports	0.03	0.33	0.40	0.34	0.10	0.25	0.32
Imports	0.13	0.45	0.36	0.32	0.40	0.29	0.50
Savings	0.53	0.68	0.38	0.39	0.33	0.51	0.37
Net exports	− 0.18	− 0.50	− 0.08	0.23	− 0.29	− 0.16	0.07
Panel B. Supply side							
Productivity	0.42	0.53	− 0.07	0.21	0.04	0.27	0.36
Capital	0.33	0.18	0.08	0.17	0.09	0.31	0.55
Labour	0.42	0.59	0.36	0.39	− 0.17	0.42	0.60

(*Source*: Author's calculations)

To turn to the *production side*, we see that US labour input has the highest correlation with its counterpart abroad, ranging from 0.60 with the UK to a low of minus 0.17 with Italy. International productivity levels also tend to be positively correlated, though not to the extent of labour input. Changes in capital formation have a low international correlation, consistent with other facets of this input documented above. The highest international business cycle correlations are between Canada and the United States, geographic neighbours with similar institutions and extensive trade relations.

Real Exchange Rates and the Terms of Trade

The two key international relative prices are the real exchange rate and the terms of trade. The real exchange rate is:

$$Q_t^R = \ln\left(E_t P_t^* / P_t\right) \tag{7}$$

where E_t is the nominal exchange rate between the home and foreign country and P_t and P_t^* are home and foreign price indices (usually the consumer price index), respectively. In words: Q_t^R is the cost of the foreign consumption basket relative to the domestic consumption basket after converting to a common currency. According to the purchasing power parity proposition, the dollar goes just as far in foreign countries as it does in the United States in terms of purchasing power. This implies that $Q_t^R = 1$ at each point in time.

In practice, however, the real exchange rate is highly variable and very persistent. High variability suggests large absolute departures from parity, while high persistence implies that, when a price gap opens up internationally, it tends to remain open for many months rather than days or weeks. In terms of the time series measurement of this property, at business cycle frequencies, it appears that the real and nominal exchange rates have approximately the same variance while the price ratio term $\left(P_t^* / P_t\right)$ is very stable. For example, the standard deviation of the nominal exchange rate between the United States and France is about 8.52, close to the standard deviation of their bilateral real exchange rate at 7.95, while the price ratio has a standard deviation of only 1.17 (see Table 3). These numbers are typical of US bilateral real exchange rates with respect to other industrialized countries. One also finds that the real exchange rate is not highly correlated with quantity variables such as output or even net exports (not shown).

To turn to the terms of trade, it is defined as:

$$Q_t^T = P_t^m / P_t^x \tag{8}$$

where P_t^m and P_t^x are import and export price indices for a particular country. Since these price indices are domestic deflators, they are already expressed in the home currency terms, and the spot exchange rate is not needed to convert them

International Real Business Cycles, Table 3 Cyclical properties of real exchange rates and the terms of trade

	US	Australia	Canada	France	Germany	Italy	Japan	UK
Panel A. Standard deviations								
Price ratio				1.17	1.42	1.67		1.74
Nominal exchange rate				8.52	8.37	8.51		8.20
Real exchange rate				7.95	8.06	7.80		
Terms of trade	2.90	5.21	2.44	3.50	2.61		5.68	2.64
Trade ratio		9.94	4.60	3.66	3.90		7.29	3.94
Panel B. Contemporaneous cross correlations								
Output and net exports	− 0.30	− 0.19	− 0.43	− 0.30	− 0.05		− 0.23	− 0.25
Output and the terms of trade	− 0.08	− 0.30	− 0.11	− 0.14	− 0.09		− 0.09	0.22
Terms of trade and net exports	0.28	− 0.07	0.06	− 0.51	0.00		− 0.50	− 0.54

Sources: Terms of trade moments are from Table 1 in Backus and Crucini (2000). Sample periods are as follows: Canada, the United Kingdom and the United States, 1955Q1–1990Q3; Australia, 1960Q1–1990Q3; France, 1970Q1–1990Q3; Germany, 1968Q1–1990Q3; Italy, 1970Q1–1990Q2; Japan, 1955Q2–1990Q3. Real exchange rate moments are from Chari et al. (2002); sample period is 1973Q1–2000Q1

to common units. Unlike the real exchange rate, economic theory does not place strong restrictions on the time series or cross-country behaviour of the terms of trade. Given the presumption that countries import different goods from those they export, we expect the terms of trade to be different from unity, and it should fluctuate, too.

Australia and Japan have the highest terms-of-trade variability, about twice that of the other countries, with the exception of France, which experiences terms-of-trade variability between these extremes. The terms of trade does not have a robust correlation with either output or net exports.

Modelling International Business Cycles

Quantitative theoretical investigations of business cycles seek to account for business cycle facts using models in which consumers are thoughtful and informed, firms employ workers and utilize capital efficiently, and policymakers use a combination of rules and discretion to achieve various economic objectives. The key dimensions of study are those unique to international economics: matching the international character of the world business cycle and the business cycle properties of the current account, the real exchange rate and the terms of trade.

The Current Account

The most rudimentary model of current account behaviour is one in which a small open economy faces an exogenous world interest rate and income stream. To fix ideas, think of a small country that produces mostly oil with perfect access to international capital markets. If the country is always producing at capacity, all of its income variation is due to changes in the price of oil in world markets. What does the intertemporal approach to the current account predict in this circumstance?

The theory reduces the NIPA identity to: $S_t = Y_t - C_t$, so that consumption decisions effectively determine saving decisions. Investment is absent since we are abstracting from changes in production capacity and its utilization. While this model seems simplistic, the identity is deceptive since it suggests that only current income enters into the current consumption–savings decision. In fact, the most widely used set-up has its roots in the seminal contribution of Friedman (1957), with individuals assumed to be able to draw upon the entire present discounted value of their future labour income. Whereas current income is the traditional argument in the Keynesian consumption function, wealth plays this role in modern macroeconomics. Since wealth is the sum of the market value of financial assets and all future anticipated flows of income, expectations play a central role in the modern consumption function. (There are many extensions to this basic framework that prevent individuals from drawing upon their lifetime wealth for present consumption: collateral requirements, limits on debt-to-income ratios and credit histories. Discussion of these extensions is beyond the scope of this survey.)

Much of the intuition for the impact of a changing income profile on the current account of a small open economy is available from Quah's (1990) formulation of the permanent-income hypothesis. He assumes a constant interest rate, quadratic preferences and rational expectations. He allows income to contain both permanent and transitory shocks. If we assume income follows a first-order autoregressive process: $Y_{t+1} = \rho Y_t + \nu_{t+1}$, where ν_{t+1} is news about income (that is, under rational expectations, news about income is: $E_{t+1} Y_{t+1} - E_t Y_{t+1} = \nu_{t+1}$), the predicted change in consumption in response to this news is:

$$\Delta C_{t+1} = \frac{r}{1 + r - \rho} \nu_{t+1} \tag{3.1}$$

and the change in the current account on impact, on the assumption it was in balance initially, is:

$$\begin{aligned} \Delta CA_{t+1} &= \Delta Y_{t+1} - \Delta C_t \\ &= \nu_{t+1} - \frac{r}{1 + r - \rho} \nu_{t+1}. \end{aligned}$$

Since output deviations from trend (the business cycle) are persistent, it is safe to assume, $\rho > 0$. A plausible value for r is 0.05 (a five per cent real interest rate). Note that the consumption response depends positively on persistence since wealth effects are rising in the persistence of the income change. As persistence moves from zero toward

unity, the effect on the current account rises from close to unity toward zero. This algebra delivers a key prediction of the intertemporal approach, that consumption smoothing leads to current account surpluses during booms unless the income change is viewed as permanent (that is, $\rho = 1$) in which case the current account is predicted to remain unchanged.

While there is evidence to suggest an interest rate channel on consumption, it does not help to resolve the counterfactual prediction of a pro-cyclical current account from the consumption side, just established. There are two reasons for this. First, if interest rates are higher during a boom in the home country, individuals would tend to tilt consumption from current to future periods (that is, postpone durable goods purchases) – the intertemporal substitution effect. This would reinforce rather than overturn our prediction that the current account moves into surplus during a boom. If real interest rates actually fell during a boom, the intertemporal substitution effect would operate in the right direction, but the evidence on the cyclicality of the real interest is ambiguous. Second, when we move to a general equilibrium setting, incorporating home and foreign responses, the increase in the real interest rate is shared by the two countries and therefore incapable of delivering the asymmetric consumption responses necessary to move the current account balance. This leaves us with the need to look elsewhere for a channel that moves the current account in a countercylical direction.

To return to the algebra of the current account identity, it would appear that what is needed for a countercyclical current account is for domestic investment to rise more than domestic savings during a business cycle expansion:

$$\begin{aligned} \Delta CA_t &= \Delta S_t - \Delta I_t \\ &= (\Delta Y_t - \Delta C_t) - \Delta I_t. \end{aligned} \tag{3.3}$$

The identity reveals the tension between the consumption smoothing channel, whereby a transitory change in income is mostly saved, pushing the current account towards surplus and the investment channel, which pulls in the opposite direction, towards a deficit.

In a model with only one good, the consumption smoothing channel wins the contest unless the shocks are highly persistent (see, for example, Backus et al. 1992; Baxter and Crucini 1993; Mendoza 1991). Persistence, by increasing the impact response of consumption due to the larger wealth effect, helps to push the current account towards balance, leaving the investment channel to produce a deficit. Extensive empirical investigations of the intertemporal approach to the current account may be found in Glick and Rogoff (1995) and Nason and Rogers (2006). (Sachs (1981) provides early evidence on the investment channel.)

Extensions of the model to multiple goods helps avoid this unpleasant arithmetic because individuals want to increase consumption of both the domestic good and the foreign good, increasing import demand and reinforcing the tendency towards a deficit from a traditional trade channel. Demonstrations of this effect under complete and incomplete risk sharing are found in Backus et al. (1994) and Arvanitis and Mikkola (1996), respectively. (JoAnne Feeney (1994) provides an insightful exposition of this issue.)

To summarize, early developments of the intertemporal approach to the current account emphasized the consumption smoothing channel and predicted that current account surpluses would occur when output was temporarily above trend. Current account surpluses are often described as *good* based on the idea that surpluses flow from good economic times. The complete model of the current account adds investment dynamics and allows for the possibility that investment-led booms produce current account deficits. These theoretical developments and their empirical implications have led to a more balanced view of the current account: that we need to understand the sources of the changes in the current account before making value judgements about them. Kollman (1998), appears to be the first quantitative simulation of US and European current account dynamics using a modern real business cycle analysis that incorporates variation in productivity, government spending and national tax rates.

The World Business Cycle

Conceptually, the world business cycle is simple to define: the deviation of world output from its growth trend. The practical difficulty is the measurement of world output because national output is denominated in domestic currency. Converting nominal output into a common currency using spot nominal exchange rates greatly exaggerates fluctuations in output because nominal exchange rates are much more volatile than either real production or price levels. Moreover, prices vary considerably across nations even after conversion to a common currency, making it difficult to construct an appropriate deflator to convert nominal gross international product into real gross international product. Here we follow much of the existing literature and use real gross domestic product of each country, and compute correlations across them. If real output is highly correlated across countries, we have evidence of a world business cycle. As we documented earlier, most macroeconomic aggregates are positively correlated across countries, indicative of a world business cycle. How do business cycle researchers account for this fact?

There are two channels through which positive economic co-movement may arise: endogenous propagation and exogenous propagation. Positive endogenous propagation refers to a situation in which a disturbance originating in one country has a positive impact on both home and foreign output levels. For example, rapid development in China drives up demand for crude petroleum and fuels economic expansions in countries that are specialized in oil production. Positive exogenous propagation refers to the correlation of shocks across countries. For example, the Second World War witnessed dramatic increases in national output in most industrialized countries as government spending rapidly expanded during the conflict. In practice, endogenous propagation and exogenous propagation are difficult to distinguish, presenting one of the key challenges of business cycle research.

Real business cycle researchers have devoted most of their effort to measuring total factor productivity, which has been found to be highly persistent and positively correlated across countries. Correlations over the business cycle are typically lower than correlations over long periods of measurement, suggesting more commonality in the technological trend than in the productivity cycle. (When analysis extends to small developing countries the business cycle correlations sometimes exceed the growth correlations.) Given the lower correlation of fiscal variables with the cycle and their modest cyclical variation, it is not surprising that they have received less attention in empirical and theoretical analysis than productivity. Two key studies of the empirical behaviour of international taxes and their equilibrium implications using dynamic equilibrium theory are Mendoza et al. (1994), and Mendoza and Tesar (1998), respectively. Both studies suggest international taxation is more relevant for secular and long-run trends than it is for business-cycle fluctuations.

Apart from the obvious role of the correlation of the shocks that drive business cycles, three economic factors have proven critical in determining the ability of dynamic equilibrium models to generate international co-movements resembling those we see in the data. The first is the extent to which domestic and foreign goods are substitutes in demand. The second is the extent to which factors of production are internationally mobile. The third is the extent of international financial linkages.

The first generation of models by Backus et al. (1992) and Baxter and Crucini (1993) followed the analytical structure of the closed economy models by Kydland and Prescott (1982) and King et al. (1988) quite closely. Despite the similarity, however, international economists were immediately confronted with two key modelling issues. The first had to do with factor mobility across countries, which is obviously absent in the closed economy setting. The mobility of labour across countries seemed minor enough to ignore, physical capital mobility was not. Since physical capital takes real resources to reallocate, the standard approach has been to subject capital accumulation to adjustment costs (or time to build as in Backus et al. 1992). Without some cost of physical capital mobility, capital would be predicted to move rapidly and in large amounts

across national boundaries in response to persistent changes in productivity or taxes. Such factor movements generate strongly negative correlations of output from the supply side and unrealistically volatile investment over the business cycle.

The second issue model builders were confronted with was asset market structure. Much of aggregative economics is predicated on the basis that idiosyncratic shocks are irrelevant to macroeconomic fluctuations. In an economy with millions of individual agents and thousands of firms, the law of large numbers combined with not-too-objectionable restrictions of preferences and technology provided a compelling argument to abstract from idiosyncratic variation. At the aggregative international level, the number of shocks is small (in many models it equals the number of countries), and countries are large and few in number. Thus, it makes little sense to rely on the law of large numbers, so researchers adopted the assumption that agents pool nation-specific risks, avoiding the need to track the wealth distribution across countries.

Unfortunately, complete risk pooling in the one-sector model leads to a presumption that output is negatively correlated across countries while consumption is close to perfectly positively correlated. In the data, the reverse rankings of correlation tend to prevail, and the absolute level of consumption correlations is well below unity. The prediction of near-perfect consumption co-movement across countries derives from the risk-pooling assumption and the fact that agents face common prices and interest rates.

The negative correlation of income is driven by cost-minimizing production decisions where firms allocate plants and equipment to the most productive location. Thus, an increase in home productivity increases domestic output relative to foreign output directly, and this is reinforced by the flow of capital from the less productive country to the more productive country. Risk pooling also enhances the supply-side response by neutralizing the wealth redistribution effects on home and foreign labour supplies.

Debate continues as to what the appropriate asset market structure should be and how to incorporate changes in asset diversification in business cycle models. Baxter and Crucini (1995) and Kehoe and Perri (2002) show that, when risk pooling is limited, positive output co-movements are more likely to arise the more persistent are the deviations to relative international productivity. Also, consumption correlations may actually fall below output correlations if the shocks are close to permanent, a feature that is prevalent in the data and difficult to explain from a number of standard theoretical paradigms.

Researchers have had more success accounting for positive international output co-movement in models where countries depend on their trading partners for final goods or intermediate inputs they themselves do not produce. Examples of work along these lines include an extension of the multisector model with intermediate inputs of Long and Plosser (1983) to the open economy by Ambler et al. (2002), a model of the North–South business cycle by Michael Kouparitsas (1996) which emphasizes trade of manufactures for primary inputs across these two regions, and the introduction of home production by Canova and Ubide (1998). A contribution that extends the incomplete markets model developed by Baxter and Crucini (1995) to the two-good setting is Arvanitis and Mikkola (1996).

Real Exchange Rates and the Terms of Trade

Multiple sectors take centre stage when one considers the real exchange rate and the terms of trade. Approaches to international relative price determination may be usefully placed into two categories. One category focuses on the determination of international relative prices of different goods. A second category focuses on deviations from the law of one price, meaning identical goods trade at different prices in different countries.

A classic contribution in the former category is Backus et al. (1994) (BKK), who develop a two-country, two-good model. Each country specializes in the production of one of the two goods and the two goods are combined in production, via an Armington aggregator, to create a composite final good which is, in effect, the single final good in each economy.

The Armington aggregator is a function that describes how substitutable the two goods are in achieving a particular output level of the final good. To match low trade shares with the specialization-in-production assumption, home bias is assumed in the aggregator function. This means that the home country uses more of the home good when producing the composite good, and the foreign country behaves symmetrically.

This is an elegant model that ties in nicely with the one-sector two-country framework. The key difference between this model and the one-sector model is that specialization provides a motivation for keeping production levels more nearly equal across locations, since individuals have demands for each type of good. The model allows us to study the terms of trade, a key international relative price absent from the one-sector model, by construction. In the BKK model, the terms of trade and trade ratio are related as follows:

$$q_t = \ln\left(\frac{p_{bt}}{p_{at}}\right) = \omega + \frac{1}{\sigma}\ln(a_t/b_t). \qquad (3.4)$$

In words: an increase in production of the home good, a, drives down its relative price. The home terms of trade turn against the country experiencing the expansion, a pro-cyclical terms of trade, as BKK define it. In the data, the correlation varies substantially across countries in magnitude and sign. The model has difficulty matching both the observed volatility of the terms of trade and the quantity ratio; as the Euler equation makes clear, there is a trade-off between terms of trade and quantity ratio variability as the elasticity is altered. If we view a and b as the final consumption levels of each good, the quantity ratio is not nearly volatile enough, given a plausible degree of elasticity, to generate the terms of trade variation we see in the data. Backus and Crucini (2000) add an oil producing region (and sector) and find that the model does better in matching the cyclicality of the terms of trade and the trade balance than the original BKK model. Kose (2002) provides an extensive quantitative analysis of the variation of international relative prices and their role in the business cycles of small open economies.

Models that consider deviations from the law of one price differ in the source of the price deviations and their duration. Sticky-price models consider the deviations to be transitory, with nominal prices responding with a lag to changes in the economic environment. These models also assume trade in an infinite number of varieties, which allows individual firms to charge a markup of price over marginal cost. Key contributions in this area are Svennson and van Wijgenbergen (1989) and Obstfeld and Rogoff (1995).

Trade cost models treat price deviations as a consequence of a real resource cost of trading, or operating businesses, in different locations. The simplest version allows prices to vary across locations by a shipping cost, usually treated as proportional to the marginal cost of the producer/supplier. The seminal contribution is Samuelson (1952), with more recent contributions by Eaton and Kortum (2002) and Sercu et al. (1995). An alternative variant is to distinguish traded and non-traded goods with traded goods not subject to trade costs and non-traded goods assumed to be subject to prohibitive trade costs, as in the original Salter (1959) and Swan (1960) models. Stockman and Tesar (1995) conduct a quantitative investigation of the business cycle predictions of this class of model.

Recent efforts have focused on quantifying the role of sticky prices, imperfect competition and trade costs in accounting for international relative price deviations and their business cycle implications. Chari et al. (2002) conduct a quantitative evaluation of the sticky-price, imperfect-competition model and find that it can account for only a small part of the persistence and somewhat more of the volatility of the real exchange rate. (See also Betts and Devereux 2000; Bergin and Feenstra 2000.) Corsetti et al. (2005) and Ravn and Mazzenga (2004) show the promise of models that combine imperfect competition with real trading costs.

What is missing from existing models is a clear distinction between economic activities that take place at the dock and exchange in retail markets. Transportation costs alone cannot account for all of the retail price dispersion we observe. Presumably, this is because much of what the retail

market entails are local inputs of land, labour and infrastructure (some of it publicly provided). Models and empirical evidence are just now being developed to make these distinctions, such as Burstein et al. (2003) and Crucini et al. (2005), respectively.

See Also

- ▶ Business Cycle Measurement
- ▶ Macroeconomic Effects of International Trade
- ▶ Real Business Cycles
- ▶ Stylized Facts.

Acknowledgments The author is grateful to Ben Eden for comments and to Hakan Yilmazkuday for excellent research assistance

Bibliography

Ambler, S., E. Cardia, and C. Zimmerman. 2002. International transmission of the business cycle in a multi-sector model. *European Economic Review* 46: 273–300.

Ambler, S., E. Cardia, and C. Zimmermann. 2004. International business cycles: What are the facts? *Journal of Monetary Economics* 51: 257–276.

Arvanitis, A.V., and A. Mikkola. 1996. Asset-market structure and international trade dynamics. *American Economic Review* 86: 67–70.

Backus, D.K., and M.J. Crucini. 2000. Oil prices and the terms of trade. *Journal of International Economics* 50: 185–213.

Backus, D.K., P. Kehoe, and F. Kydland. 1992. International real business cycles. *Journal of Political Economy* 100: 745–775.

Backus, D.K., P. Kehoe, and F. Kydland. 1994. Dynamics of the trade balance and the terms of trade: The J-curve. *American Economic Review* 84: 84–103.

Baxter, M., and M.J. Crucini. 1993. Explaining saving-investment correlations. *American Economic Review* 83: 416–436.

Baxter, M., and M.J. Crucini. 1995. Business cycles and the asset structure of foreign trade. *International Economic Review* 36: 821–854.

Baxter, M., and D. Farr. 2005. Variable capital utilization and international business cycles. *Journal of International Economics* 65: 335–347.

Bergin, P.R., and R.C. Feenstra. 2000. Staggered price setting, translog preferences, and endogenous persistence. *Journal of Monetary Economics* 45: 657–680.

Betts, C., and M. Devereux. 2000. Exchange rate dynamics in a model of pricing to market. *Journal of International Economics* 50: 215–244.

Beveridge, S., and C. Nelson. 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics* 7: 151–174.

Burns, A., and W. Mitchell. 1946. *Measuring business cycles*. New York: NBER.

Burstein, A.T., J.C. Neves, and S. Rebelo. 2003. Distribution costs and real exchange rate dynamics during exchange-rate based stabilizations. *Journal of Monetary Economics* 50: 1189–1214.

Canova, F., and A.J. Ubide. 1998. International business cycles, financial markets and household production. *Journal of Economic Dynamics and Control* 22: 545–572.

Chari, V.V., P.J. Kehoe, and E.R. McGrattan. 2002. Can sticky price models generate volatile and persistent real exchange rates? *Review of Economic Studies* 69: 533–563.

Cochrane, J.H. 1994. Permanent and transitory components of GNP and stock prices. *Quarterly Journal of Economics* 109: 241–265.

Corsetti, G., Dedola, L. and Leduc, S. 2005. International risk-sharing and the transmission of productivity shocks. International Finance Discussion Paper No. 826. Board of Governors of the Federal Reserve Board.

Crucini, M.J., and M. Shintani. 2006. *International comovement: Is theory ahead of business cycle measurement?* Mimeo: Vanderbilt University.

Crucini, M.J., C.I. Telmer, and M. Zachariadis. 2005. Understanding European real exchange rates. *American Economic Review* 95: 724–738.

Eaton, J., and S. Kortum. 2002. Technology, geography, and trade. *Econometrica* 70: 1741–1779.

Feeney, J. 1994. Goods and asset market interdependence in a risky world. *International Economic Review* 35: 551–563.

Feldstein, M., and C.I. Horioka. 1980. Domestic saving and international capital flows. *Economic Journal* 90: 314–329.

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

Glick, R., and K. Rogoff. 1995. Global versus country-specific productivity shocks and the current account. *Journal of Monetary Economics* 35: 159–192.

Hodrick, R., and E.C. Prescott. 1997. Post-war U.S. business cycles: An empirical investigation. *Journal of Money. Credit and Banking* 29: 1–16.

Imbs, J. 1994. Technology, growth and the business cycle. *Journal of Monetary Economics* 44: 65–80.

Kehoe, P.J., and F. Perri. 2002. International business cycles with endogenous incomplete markets. *Econometrica* 70: 907–928.

King, R.G., C.I. Plosser, and S. Rebelo. 1988. Production, growth, and business cycles I: The basic neoclassical model. *Journal of Monetary Economics* 21: 195–232.
Kollman, R. 1998. U.S. trade balance dynamics: The role of fiscal policy and productivity shocks and of financial market linkages. *Journal of International Money and Finance* 17: 637–669.
Kose, A. 2002. Explaining business cycles in small open economies: How much do world prices matter? *Journal of International Economics* 56: 299–327.
Kose, A., C. Otrok, and C. Whiteman. 2003. International business cycles: World, region, and country-specific factors. *American Economic Review* 93: 1216–1239.
Kouparitsas, M. 1996. North–South business cycles. Working Paper No. 96–9, Federal reserve bank of chicago.
Kydland, F.E., and E.C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.
Long, J., and C.I. Plosser. 1983. Real business cycles. *Journal of Political Economy* 91: 39–69.
Mendoza, E. 1991. Real business cycles in a small open economy. *American Economic Review* 81: 797–818.
Mendoza, E., A. Razin, and L. Tesar. 1994. Effective tax rates in macroeconomics: Cross-country estimates of tax rates on factor incomes and consumption. *Journal of Monetary Economics* 34: 297–323.
Mendoza, E., and L. Tesar. 1998. The international ramifications of tax reforms: Supply-side economics in a global economy. *American Economic Review* 88: 226–245.
Nason, J., and J. Rogers. 2006. The present-value model of the current account has been rejected: Round up the usual suspects. *Journal of International Economics* 68: 159–187.
Obstfeld, M., and K. Rogoff. 1995. Exchange rate dynamics redux. *Journal of Political Economy* 102: 624–660.
Quah, D. 1990. Permanent and transitory movements in labor income: An explanation for 'excess smoothness' in consumption. *Journal of Political Economy* 98: 449–475.
Ravn, M., and E. Mazzenga. 2004. International business cycles: The quantitative role of transportation costs. *Journal of International Money and Finance* 23: 645–671.
Sachs, J.D. 1981. The current account and macroeconomic adjustment in the 1970s. *Brookings Papers on Economic Activity* 1981(1): 201–268.
Salter, W.E. 1959. Internal and external balance: The role of price and expenditure effects. *Economic Record* 35: 226–238.
Samuelson, P.A. 1952. Spatial price equilibrium and linear programming. *American Economic Review* 42: 283–303.
Sercu, P., R. Uppal, and C. van Hulle. 1995. The exchange rate in the presence of transaction costs: Implications for tests of purchasing power parity. *Journal of Finance* 50: 1309–1319.
Stock, J., and M. Watson. 2005. Understanding changes in international business cycle dynamics. *Journal of the European Economic Association* 3: 968–1006.
Stockman, A., and L. Tesar. 1995. Tastes and technology in a two-country model of the business cycle: Explaining international comovements. *American Economic Review* 85: 168–185.
Svennson, L., and S. van Wijgenbergen. 1989. Excess capacity, monopolistic competition and international transmission of monetary disturbances. *Economic Journal* 99: 785–805.
Swan, T. 1960. Economic control in a dependent economy. *Economic Record* 36: 51–66.

International Reserves

Joshua Aizenman

I

Abstract

Developing countries, particularly in East Asia, account for most of the large increase in international reserves–GDP ratios in recent decades. Possible explanations include self-insurance against the output costs of sudden stops; precautionary fiscal demand by countries with inelastic fiscal outlays, sovereign risk, volatile and limited tax capacity; and a modern incarnation of mercantilism. Empirical studies reveal that the 1997–8 East Asian financial crisis triggered a sharp increase in hoarding international reserves. They suggest prominent roles for the precautionary demand and self-insurance motives and conclude that the financial integration of developing countries is associated with greater hoarding of international reserves.

Keywords

Asian miracle; Buffer stock model; Exchange-rate flexibility; Hot money; International capital flows; International reserves; Liquidity crises; Option pricing theory; Self-insurance

JEL Classifications
F3

International reserves are the liquid external assets under the control of the central bank. An intriguing development since the 1960s has been that, despite the proliferation of greater exchange rate flexibility, international reserves–GDP ratios increased substantially. Flood and Marion (2002) report that reserve holdings have trended upwards; at the end of 1999, reserves were about 6 per cent of global GDP, 3.5 times what they were at the end of 1960 and 50 per cent higher than in 1990. Practically all the increase in reserves-GDP holding has been by developing countries, mostly concentrated in East Asia.

These developments stirred lively debate among economists and financial observers. The earlier literature focused on using international reserves as a buffer stock, part of the management of an adjustable-peg or managed-floating exchange- rate regime. Accordingly, optimal reserves balance the macroeconomic adjustment costs incurred in the absence of reserves with the opportunity cost of holding reserves (see Frenkel and Jovanovic 1981). The buffer stock model predicts that average reserves depend negatively on adjustment costs, the opportunity cost of reserves, and exchange rate flexibility; and positively on GDP and on reserve volatility, driven frequently by the underlying volatility of international trade. Overall, the literature of the 1980s supported these predictions; see Frenkel (1983), Edwards (1983), and Flood and Marion (2002) for a recent review.

While useful, the buffer stock model has limited capacity to account for the recent development in hoarding international reserves – the greater flexibility of the exchange rates exhibited in recent decades should work in the direction of reducing reserve hoarding, in contrast to the trends reported above. As an indication of excess hoarding, observers noted that developing countries frequently borrow at much higher interest rates than the one paid on reserves.

The recent literature provided several interpretations for these puzzles, focusing on the observation that the deeper financial integration of developing countries has increased exposure to volatile short-term inflows of capital (dubbed 'hot money'), subject to frequent sudden stops and reversals (see Calvo 1998; Edwards 2004). Looking at the 1980s and 1990s, Aizenman and Marion (2003a) pointed out that the magnitude and speed of the reversal of capital flows throughout the 1997–8 crisis surprised most observers. Most viewed East Asian countries as being less vulnerable to the perils associated with hot money than Latin American countries. After all, East Asian countries were more open to international trade, had sounder fiscal policies, and much stronger growth performance. In retrospect, the 1997–8 crisis exposed hidden vulnerabilities of East Asian countries, forcing the market to update the probability of sudden stops affecting all countries.

The above observations suggest that hoarding international reserves can be viewed as a precautionary adjustment, reflecting the desire for self-insurance against exposure to future sudden stops. Self-insurance has several interpretations. The first focuses on precautionary hoarding of international reserves needed to stabilize fiscal expenditure in developing countries (see Aizenman and Marion 2003b). Specifically, a country characterized by volatile output, inelastic demand for fiscal outlays, high tax collection costs and sovereign risk may want to accumulate both international reserves and external debt. External debt allows the country to smooth consumption when output is volatile. International reserves that are beyond the reach of creditors would allow such a country to smooth consumption in the event that adverse shocks trigger a default on foreign debt. Political instability, by taxing the effective return on reserves, can reduce desired current reserve holdings. The tests reported by Aizenman and Marion (2003b) are consistent with this interpretation. Another version of self-insurance and precautionary demand for international reserves follows the earlier work of Ben-Bassat and Gottlieb (1992), viewing international reserves as output stabilizers (see Aizenman and Lee 2005; see Lee 2004, for insurance perspectives of international reserves applying the option pricing theory). Accordingly, international reserves can reduce the probability of an output drop induced by a

sudden stop and/or the depth of the output collapse when the sudden stop materializes (see Kaminsky and Reinhart 1999).

The views linking the large increase in hoarding reserves to deeper financial integration face a well-known contender in a modern incarnation of mercantilism: international reserves accumulations triggered by concerns about export competitiveness. This explanation has been advanced by Dooley et al. (2003), especially in the context of China. They interpret reserves accumulation as a by-product of promoting exports, which is needed to create better jobs, thereby absorbing abundant labour in traditional sectors, mostly in agriculture. While intellectually intriguing, this interpretation remains debatable. Some have pointed out that high export growth is not the new kid on the block – it is the story of East Asia since the 1950s. Yet the large increase in hoarding reserves happened mostly after 1997. This issue is of more than academic importance: the precautionary approach links reserves accumulation directly to exposure to sudden stops, capital flight and volatility, whereas the mercantilist approach views reserves accumulation as a residual of an industrial policy, a policy that may impose negative externalities on other trading partners.

Aizenman and Lee (2005) test the importance of precautionary and mercantilist motives in accounting for the hoarding of international reserves by developing countries. While variables associated with the mercantilist motive (like lagged export growth and deviation from purchasing power parity) are statistically significant, their economic importance in accounting for reserve hoarding is close to zero and is dwarfed by other variables. Overall, the empirical results are in line with the precautionary demand. The effects of financial crises have been localized, increasing reserve hoarding in the aftermath of crises mostly in countries located in the affected region, but not in other regions. A more liberal capital account regime is found to increase the amount of international reserves, in line with the precautionary view. These results, however, do not imply that the hoarding of reserves by countries is optimal or efficient. Making inferences regarding efficiency would require having a detailed model and much more information, including an assessment of the probability and output costs of sudden stops, and the opportunity cost of reserves. To conclude, greater exposure of developing countries to sudden stops and reversals of hot money as well as growing trade openness go a long way towards accounting for the observed increase in international reserves–GDP ratios by developing markets.

See Also

- ▶ Exchange Rate Volatility
- ▶ International Capital Flows
- ▶ Liquidity Constraints

Bibliography

Aizenman, J. and J. Lee. 2005. *International reserves: Precautionary versus mercantilist views, theory and evidence*. Working paper no. 11366. Cambridge, MA: NBER.

Aizenman, J., and N. Marion. 2003a. The high demand for international reserves in the Far East: What's going on? *Journal of the Japanese and International Economies* 17: 370–400.

Aizenman, J., and N. Marion. 2003b. International reserves holdings with sovereign risk and costly tax collection. *Economic Journal* 114: 569–591.

Ben-Bassat, A., and D. Gottlieb. 1992. Optimal international reserves and sovereign risk. *Journal of International Economics* 33: 345–362.

Calvo, G. 1998. Capital flows and capital-market crises: The simple economics of sudden stops. *Journal of Applied Economics* 1: 35–54.

Dooley, M., D. Folkerts-Landau, and P. Garber. 2003. *An essay on the revived Bretton Woods system*. Working paper no. 9971. Cambridge, MA: NBER.

Edwards, S. 1983. The demand for international reserves and exchange rate adjustments: The case of LDCs, 1964–1972. *Economica* 50: 269–280.

Edwards, S. 2004. Thirty years of current account imbalances, current account reversals, and sudden stops. *IMF Staff Papers* 51(Special Issue): 1–49.

Flood, R., and P. Marion. 2002. Holding international reserves in an era of high capital mobility. In *Brookings trade forum 2001*, ed. S. Collins and D. Rodrik. Washington, DC: Brookings Institution Press.

Frenkel, J. 1983. International liquidity and monetary control. In *International money and credit: The policy roles*, ed. G. von Furstenberg. Washington, DC: International Monetary Fund.

I

Frenkel, J., and B. Jovanovic. 1981. Optimal international reserves: A stochastic framework. *Economic Journal* 91: 507–514.

Kaminsky, G., and C. Reinhart. 1999. The twin crises. The causes of banking and balance-of-payments problems. *American Economic Review* 89: 473–500.

Lee, J. 2004. *Insurance value of international reserves*. Working paper 04/175. Washington, DC: IMF.

International Trade

John S. Chipman

Edgeworth (1894) opened his survey of the theory of international values with the provocative statement: 'International trade meaning in plain English trade between nations, it is not surprising that the term should mean something else in Political Economy'. This could equally well be said today. What distinguishes international from domestic trade is the greater prevalence of barriers (both natural and artificial) to trade and factor movements in the former; different currencies; and (perhaps most important) autonomous governments, leading to a pattern of shocks which impact different countries in different ways. Because of these differences, a different type of theoretical model is called for. For example, international immobility of factors results in greater disparity in relative factor endowments among countries than among regions of the same country; these disparities may make it reasonable, as a first approximation, to ignore variations in supplies of factor services that come about in response to changes in factor rentals and commodity prices, if these variations are small in comparison with the differences in endowments. Likewise, great differences among resource endowments and productive techniques may make it reasonable to disregard differences in consumers' tastes within and across countries, even though this might be a very inappropriate type of simplification for purposes of analysing domestic trade.

The fact that national governments act independently leads to the need to analyse the effects of country-specific shocks, which take the form of intensification or liberalization of restrictions on trade or capital movements, unilateral transfers such as reparation payments, gifts, or loans, and disparities in monetary and fiscal policies. For this reason the emphasis in international-trade theory has from the beginning (Mill 1848; Marshall 1879) been on comparative statics: one wants to ascertain the qualitative, if not the quantitative, effect of a tariff or quota or transfer on the various quantities involved. To obtain unambiguous qualitative results one needs fairly drastic simplifications and strong assumptions. On the other hand, the emphasis in general-equilibrium theory (Walras 1874; Pareto 1896–97; Debreu 1959) has been on proving the existence, stability, and Pareto-optimality of competitive equilibrium, for which much milder assumptions are required. A good definition of international-trade theory as it has evolved would therefore be: 'general-equilibrium theory with structure'.

The requirements of 'simplicity' in a theory are not absolute, but vary with the goals of the theory and the technical resources available to researchers at the time. There is not much virtue in simplicity if a result that holds in a model of two countries, two commodities, and two factors does not generalize in any meaningful way to higher dimensions. With the increasing possibilities of handling large-scale models and data sets and estimating their parameters numerically, it is natural to expect a movement of both general-equilibrium traditions towards each other.

Attention will be focused here on the neoclassical model developed by Haberler (1930, 1933), Lerner (1932, 1933, 1934), Ohlin (1928, 1933), Stolper and Samuelson (1941), Samuelson (1953), and Rybczynski (1955), which Baldwin (1982) has described as the 'Haberler–Lerner–Samuelson model' – an appellation which is more accurate than the usual 'Heckscher–Ohlin theory', since the model commonly employed makes the simplifying assumption – rejected by Ohlin (1933, ch. VII) except in his illustrative Appendix I – that factors of production are inelastic in supply and indifferent among alternative occupations, allowing one to define unambiguously a country's production-possibility frontier. This model has in recent years come to

lose some of its hold on the profession – just as the Ricardian theory had in the 1930s – in favour of models that stress imperfect competition (see, e.g. Helpman and Krugman 1985). However, these latter models have so far not been successfully formulated as general-equilibrium models, and are thus still in a formative stage. It goes without saying that, in the nature of the case, a partial-equilibrium model is incapable of explaining or predicting trade patterns or analysing the effect on prices and resource allocation of trade restrictions and transfers.

The material that follows is divided into two parts. Part 1 covers the mathematical foundations of the received theory, and deals with the duality between production functions and cost functions, the concept of a national-product function, the Stolper–Samuelson and Rybczynski relations between factor rentals and commodity prices and between commodity outputs and factor endowments, the concepts of trade-demand functions and trade-utility functions, world equilibrium and its dynamic stability. Part 2 covers the applications of these basic concepts to the most noteworthy problems that have been the object of attention in the theory of international trade since its beginnings: the explanation of trade flows, the effect of unilateral transfers on sectoral prices and resource allocation, and the effect of trade restrictions such as tariffs and quotas. The reader who is interested in substantive questions is advised to proceed directly to Part 2.

Part 1. The Mathematical Foundations

Duality of Cost Functions and Production Functions

Let an industry produce a positive amount y of output of a particular product, with the aid of non-negative amounts υ_j of m primary factors of production, determining the vector $\upsilon = (\upsilon_1, \upsilon_2, \ldots, \upsilon_m)$. A *production function* f is defined over the non-negative orthant E_m^+ of m-dimensional Euclidean space, with values $y = f(\upsilon)$ on the non-negative real line E_1^+. We assume that f has the following properties:

(a) *Upper semi-continuity*: for each y the set

$$A(y) = \{\upsilon : f(\upsilon) > \} \tag{1}$$

is closed;

(b) *Quasi-concavity*: for each y, the set $A(y)$ defined by (1) is convex;

(c) *Monotonicity*: if $\upsilon, \upsilon' \in E_m^+$ are such that $\upsilon' \geqq \upsilon$, then $f(\upsilon') \geqq f(\upsilon)$.

Further properties of f will be specified later on.

We shall denote by $w = (w_1, w_2, \ldots, w_m)$ a vector of *factor rentals*, i.e. prices of the services of the m factors of production. The following conventional notation will be adhered to:

$$w \geqq 0 \text{ means } w_i \geqq 0 \text{ for all } i = 1, 2, \ldots, m;$$
$$w \geqq 0 \text{ means } w_i \geqq 0 \text{ for all } i = 1, 2, \ldots, m;$$
$$\text{and } w_i > 0 \text{ for some } i;$$

For each $y > 0$ and all $w \geq 0$ we define the *minimum total cost function* G by

$$G(w, y) = \min_{\upsilon} \{w \cdot \upsilon : f(\upsilon) \geqq y\}, \tag{2}$$

where $w \cdot \upsilon$ denotes the inner product

$$\sum_{j=1}^{m} w_j \upsilon_j$$

Mathematically, for each fixed y the function $G(\cdot, y)$ is the *support function* of the convex set $A(y)$ (cf. Fenchel 1953). It has the following properties:

(a^*) *Continuity* in w: for each y, $G(w, y)$ is continuous;

(b^*) *Concavity* in w: if $0 < \theta < 1$ then

$$(1-\theta)G(w^o, y) + \theta G(W^1, y) \leqq G\left[(1-\theta)w^o + \theta w^1, y\right];$$

(c^*) *Monotonicity*: $y' \geqq y$ implies $G(w, y') \geqq G(w, y)$ and $w' \geqq w \geq 0$ implies $G(w', y) \geqq G(w, y)$;

(d^*) *positive homogeneity* in w : $G(\lambda w, y) = \lambda G(w, y)$ for all $\lambda > 0$.

Property (a*) follows from (a) and the definition of G; property (c*) follows the definition of G and the fact that $y' \geqq y$ implies $A(y') \subseteq A(y)$; property (d*) follows immediately from the definition of G. To prove (b*), let w^0, $w^1 \geq 0$ and denote $w^\theta = (1–\theta)w^0 + \theta w^1$; from the definitions of G and $A(y)$ in (2) and (1), we have

$$G(w^o, y) \leqq w^o \cdot \upsilon \text{ for all } v \in A(y);$$
$$G(w^1, y) \leqq w^1 \cdot y \text{ for all } v \in A(y);$$

consequently,

$$(1-\theta)G(w^o, y) + \theta G\left(W^1, y\right) \leqq w^o \cdot v$$
$$\text{for all } v \in A(y).$$

Hence, in particular,

$$(1-\theta)G(w^o, y) + \theta G(w^1 \cdot y)$$
$$\leqq \min\left\{w^\theta \cdot \upsilon : \upsilon \in A(y)\right\} \equiv G\left(W^\theta, y\right),$$

which is the result sought (cf. Uzawa 1964b).

Of fundamental importance in international trade theory is the following *duality* theorem first proved by Shephard (1953). The formulation and proof contained in Theorem 1 to follow are due to Uzawa (1964b).

Theorem 1 (*Duality Theorem*). Define the set

$$B(y) = \{\upsilon : (\forall w \geq 0) w \cdot \upsilon \geqq G(w, y)\}, \qquad (3)$$

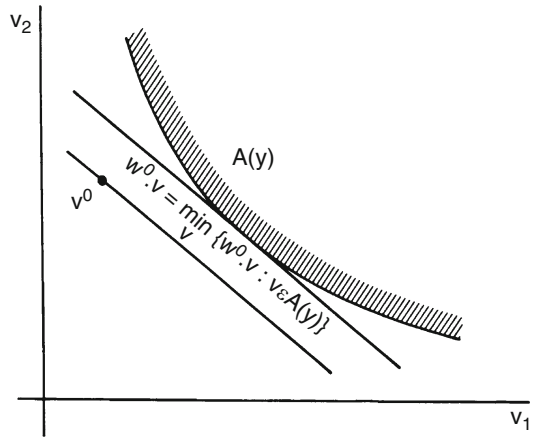
where G is defined by (2) and f satisfies properties (a), (b), (c). Then $B(y) = A(y)$, where $A(y)$ is defined by (1).

Proof: Let $\upsilon^0 \in A(y)$; then $f(\upsilon^0) \geqq y$, so for all $w \geq 0$,

$$w \cdot \upsilon^o \geqq \min_{\upsilon} \{w \cdot^u : f(\upsilon) \geqq y\} \equiv G(w, y)$$

that is, $\upsilon^0 \in B(y)$.

Conversely, suppose $\upsilon^0 \notin A(y)$. Since $A(y)$ is closed and convex by properties (a) and (b) of f, it follows from the separating hyperplane theorem of closed convex sets (cf. Fenchel 1953, p. 48) that there exists a vector $w^0 \neq 0$ such that



International Trade, Fig. 1

$$w^0 \cdot \upsilon^0 < \min_{\upsilon} \left\{w^0 . \upsilon : \upsilon \in A(y)\right\} \qquad (4)$$

(see Fig. 1). Now if w^0 has a negative component, it follows from property (c) that the corresponding component of $\upsilon \in A(y)$ may be chosen to be arbitrarily large, hence no minimum of $w^0 \cdot v$ over $A(y)$ exists; consequently, $w \geq 0$. But then the expression on the right of the inequality sign in (4) is just $G(w^0, y)$. From the definition of $B(y)$ in (3), it follows that $\upsilon^0 \notin$

$B(y)$. q.e.d.

The duality theorem may be stated in words as follows: given the function G, the set $A(y)$ may be identified with the set of all factor combinations υ which, at each constellation $w \geq 0$ of factor rentals, are at least as expensive as the minimal total cost of producing output y at factor rentals w.

Let us now explore the consequences of imposing a further condition on the production function f:

(d) *Positive homogeneity*: for all $\lambda > 0$, $f(\lambda\upsilon) = \lambda f(\upsilon)$.

From the definition of G in (2), we now have

$$G(w, y) = \min_{\upsilon} \left\{w \cdot \upsilon : f\left(\frac{\upsilon}{y}\right) \geqq 1\right\}$$
$$= \min_{b} \{yw \cdot b : f(b) \geqq 1\} \left(b = \frac{u}{y}\right)$$
$$= y \cdot \min_{b} \{w \cdot b : f(b) \geqq 1\}.$$

Thus, $G(w, y)$ factors into two terms, of which the second depends only on $w \geq 0$ and may be denoted

$$g(w) = \min_{v} \{w \cdot v : f(v) \geqq 1\}. \tag{5}$$

We therefore have

Theorem 2. If f satisfies properties (a), (b), (c), (d), then the function G of (3) factors into

$$G(w, y) = yg(w) \tag{6}$$

where g is defined by (5) and is continuous, concave, monotone, and positively homogeneous of first degree.

The properties of g specified in Theorem 2 follow directly from those of the function G.

We may now state a special form of the duality theorem for the case of homogeneous production functions.

Theorem 3. Let g be defined by (5) where f satisfies properties (a), (b), (c), (d), and let the function f^* be defined by

$$f^*(v) = \min_{v} \{w \cdot v : g(w) \geqq 1\}. \tag{7}$$

Then $f^* = f$.

Proof: Define the set

$$C(y) = \{v : [\forall w \in A^*(1)] w \cdot v \geqq y\} \tag{8}$$

where for convenience we define

$$A^*(p) = \{w : g(w) \geqq p\}. \tag{9}$$

(Since g is defined only for $w \geq 0$, $w \in A^*(p)$ implies $w \geq 0$). First we shall show that $C(y) = B(y)$, where $B(y)$ is defined by (3). From (3) and (6), if $v^0 \in B(y)$ then for all $w \in A^*(1)$, $w \cdot v^0 \geqq G(w, y) = yg(w) \geq y$, so $B(y) \subseteq C(y)$. Conversely suppose $v^0 \in C(y)$ and take any $w^0 \geq 0$. Then from the homogeneity of g we have $g[w^0/g(w^0)] = 1$, hence from the definition (8) of $C(y)$ it follows that

$$\frac{w^o}{g(w^o)} \cdot v^o \geqq y,$$

i.e., $w^0 \cdot v^0 \geqq yg(w^0)$; thus $v^0 \in B(y)$. Therefore $B(y) = C(y)$ and by Theorem 1, $C(y) = A(y)$.

Now denote $r = w/g(w)$ and consider the set

$$C'(y) = \left[v : \min_{r} \{r \cdot v : r \in A^*(1)\} \geqq y\right]. \tag{10}$$

If $r \cdot v \geqq y$ for all $r \in A^*(1)$, then a fortiori $r \cdot v \cdot \geqq y$ for the $r \in A^*(1)$ which minimizes $r \cdot v$; hence $C(y) \subseteq C'(y)$. Conversely, for all $r \in A^*(1)$ we have $r \cdot v \geqq \min_r \{r \cdot v : r \in A^*(1)\}$, so $C'(y) \subseteq C(y)$. Thus $C'(y) = C(y) = A(y)$. But from (7), (9) and (10) we have

$$C'(y) = \{v : f^*(v) \geqq y\}. \tag{11}$$

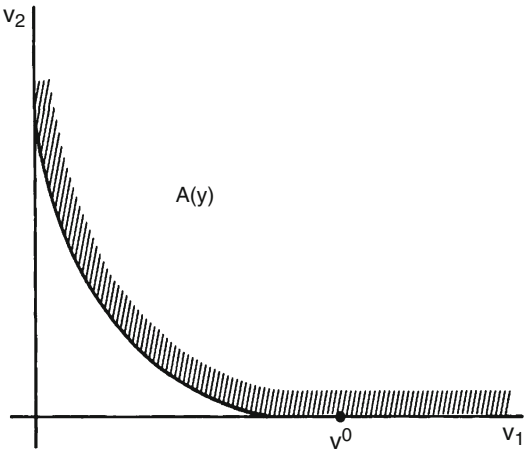
Since $A(y) = C'(y)$ for all y, therefore f and f^* coincide.

q.e.d.

Let us consider the consequences of adding to the properties (a), (b), (c) of f given in §1.1 the following further properties:

(b_1) *Strict quasi-concavity:* for each y, the set $A(y)$ defined by (1) is strictly convex;

(e) *Differentiability:* f has continuous first-order partial derivatives (Fig. 2).



International Trade, Fig. 2

For the time being, property (d) of §1.2 will not be used, but will be introduced again later on.

The problem of deriving the minimum total cost function $G(w, y)$ may be posed in terms of the following non-linear programming problem:

$$\text{minimize} \sum_{j=I}^{m} w_j \upsilon_j \text{ subject to } f(\upsilon) \geqq y, u \geqq 0 \quad (12)$$

Form the Lagrangean function

$$L(p^*, u; y, w) = \sum_{j=I}^{m} w_j \upsilon_j - p^*[f(\upsilon) - y] \quad (13)$$

where y, w are parameters and p^* is a Lagrangean multiplier. In accordance with the Kuhn–Tucker theorem (cf. Kuhn and Tucker 1951, p. 486) in order for $\upsilon^0 = (\upsilon_1^0, \upsilon_2^0, \ldots, \upsilon_m^0)$ to be a solution of the minimum problem (12), it is necessary and sufficient that υ^0 and some $p^* \geqq 0$ satisfy

$$\left.\frac{\partial L}{\partial \upsilon_j}\right|_{\upsilon=\upsilon^0} = wj - p^* \left.\frac{\partial f}{\partial v_j}\right|_{\upsilon=\upsilon^0} \geqq 0 : uj \frac{\partial L}{\partial \upsilon_j} = 0 \quad (14)$$

and

$$\sum_{j=1}^{m} \upsilon_j^0 \left.\frac{\partial L}{\partial \upsilon_j}\right|_{\upsilon j=\upsilon_j^0} = \sum_{j=1}^{m} \upsilon_j^0 \left(w_j - p^* \left.\frac{\partial L}{\partial \upsilon_j}\right|_{\upsilon j=\upsilon_j^0} \right) = 0 \quad (15)$$

as well as

$$\frac{\partial L}{\partial p^*} = -\left[f\left(\upsilon^0\right) - y\right] \geqq 0; p^* \frac{\partial L}{\partial p^*} = 0. \quad (16)$$

In the above we have used (e), but so far property (b_1) has not yet been used: Let us introduce the further properties:

(f) *indispensability*: $f(0) = 0$.
(f_1) *strict indispensability:* if υ has a component $\upsilon_j = 0$ then $f(\upsilon) = 0$.

Now suppose the solution υ^0 to (12) is such that $f(\upsilon)^0 > y$ (see Fig. 2). This violates (b_1), since strict quasi-concavity requires that if υ^0, $\upsilon^1 \in A(y)$ and $0 < \theta < 1$, the point $\upsilon^0 = (1 - \theta)\upsilon^0 + \theta\upsilon^1$ should be in the interior of $A(y)$. Suppose, however, that property (b_1) is not assumed, and that $f(\upsilon^0) > y > 0$; then $p^* = 0$ from (16) hence $w \cdot \upsilon^0 = 0$ from (15), and since $w \geq 0$ this implies that υ^0 has a zero component. Thus, if (f_1) is assumed, we have $0 = f(\upsilon^0) > y > 0$ – a contradiction. Thus, either (b_1) or (f_1) is sufficient – in conjunction with (a), (c), (e), to guarantee $f(\upsilon^0) = y$. If $w > 0$, a similar argument shows that (f) implies $f(\upsilon^0) = y$.

Now suppose that υ^0 is such that strict inequality holds in (14) for some j. Then $\upsilon^0_j = 0$ from (15) If (f_1) holds this would lead to a contradiction, since then $0 = f(\upsilon^0) \geqq y > 0$. If ($f_1$) is not assumed, but if (b_1) holds, then strict inequality in (14) implies that υ^0 has a zero component, so υ^0 is on the boundary of $A(y)$; but $2\upsilon^0$ is also on the boundary of $A(y)$, by property (c), and consequently the mid-point $1\frac{1}{2}\upsilon^0$ is as well, contradicting (b_1). Thus, if (a), (b), (c), (e) hold, then either (b_1) or (f_1) implies that equality holds in (14) for all $j = 1, 2, \ldots, m$.

Consider a solution υ^0 to (12) corresponding to a w^0 which has some zero components. Let $J = \{j: w^0_j = 0\}$. Then if $w^0 \cdot \upsilon^0 = C^0$, certainly $A(y) \subseteq \{\upsilon: w^0 \cdot \upsilon^0 \geqq C^0\}$. Let υ^1 be such that $\upsilon^1_j > \upsilon^0_j$ for $j \in J$ and $\upsilon^1_j = \upsilon^0_j$ for $j \notin J$. Then $w^0 \cdot \upsilon^1 = w^0 \cdot \upsilon^0$, hence $\upsilon^1 \in \{\upsilon: w^0 \cdot \upsilon = C^0\}$. But by condition (c), $\upsilon^1 \in A(y)$; thus υ^1 and υ^0 are both on the boundary of $A(y)$, as is $(1 - \theta)\upsilon^0 + \theta\upsilon^1$ for $0 < \theta < 1$ (see Fig. 3). This contradicts (b_1). Therefore under (b_1), a solution to (12) exists only if $w > 0$.

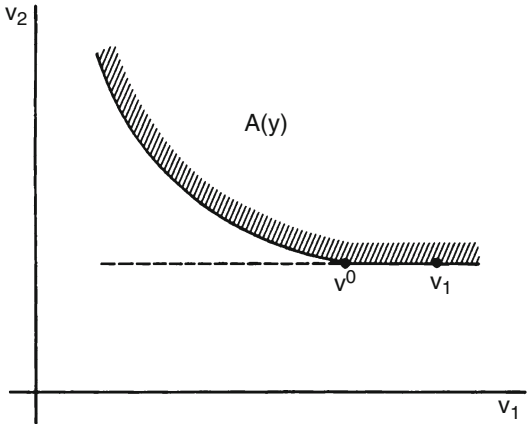
It should be noted that even if the function $G(w, y)$ of (2) is well defined in the sense

$$G(w, y) = \inf_{v} \{w \cdot \upsilon : f(\upsilon) \geqq y\}, \quad (17)$$

a solution of (12) need not exist. For example, if

$$f(\upsilon_1, \upsilon_2) = \frac{1}{\frac{1}{\upsilon_1} + \frac{1}{\upsilon_2}}$$

Then



International Trade, Fig. 3

$$G(0, w_2; y) = yw_2$$

but the *infimum* is achieved as $(\upsilon_1, \upsilon_2) \to (\infty, y)$. On the other hand a solution to (12) always exists if $w > 0$; for, choosing any $\upsilon^0 \in \text{int } A(y)$ and $w^0 > 0$, the set

$$A(y) \cap \left\{\upsilon : w^0 \cdot \upsilon \leqq w^0 \upsilon^0, \upsilon \geqq 0\right\}$$

is compact by virtue of condition (a), and from (b) and (c) the minimum of $w^0 \cdot \upsilon$ over this set is the minimum over $A(y)$.

An immediate consequence of (b_1) is that if (12) has a solution, it is *unique*. Since (12) need not have a solution unless $w > 0$, it is of some advantage to replace (b_1) by a weaker condition which still ensures uniqueness provided $w > 0$. Such a condition is

(b_2) if $\upsilon^0 \neq \upsilon^1$ and neither $\upsilon^0 \geq \upsilon^1$ nor $\upsilon^1 \geq \upsilon^0$, and if $0 < \theta < 1$, then $f[(1 - \theta)\upsilon^0 + \theta\upsilon^1] > \min[f(\upsilon^0), f(\upsilon^1)]$.

The above discussion may now be summarized in the following theorem.

Theorem 4. Let conditions (a), (b), (c), (e), (f) hold. Then if either (b_1) or (f_1) holds, any solution υ^0 to (12) has the property (Fig. 3)

$$w_j = p^* \frac{\partial f}{\partial \upsilon_j}\bigg|_{\upsilon=\upsilon^0} \quad (j = 1, 2, \ldots, m); \qquad f(\upsilon^0) = y. \tag{18}$$

If (b_1) holds, this solution is unique. If (b_2) holds and if $w > 0$, then a unique solution to (12) exists, and it satisfies (18).

We now proceed with an analysis of the solution υ of the programming problem (12) regarded as a function of the parameters $y > 0$, $w > 0$, when conditions (a), (b_2), (c), (e), (f) are assumed to hold.

In accordance with Theorem 4, the solution satisfies (18) and is unique, given y and w. Thus we have the functions

$$\upsilon_j = \upsilon(w, y) \qquad (j = 1, 2, \ldots, m). \tag{19}$$

It is shown in Fenchel (1953, pp. 102–4) that these functions are differentiable. Substituting (19) into (18) we obtain

$$p^* = w_j / \frac{\partial}{\partial \upsilon_j} f[\tilde{\upsilon}_1(w, y), \quad \tilde{\upsilon}_2(w, y), \ldots \tilde{\upsilon}_m(w, y)] = \tilde{p}^*(w, y).$$

The system of equations (18) defines a mapping $\mathscr{F}$ from the non-negative orthant of $(m + 1)$-dimensional space into itself:

$$\mathscr{F}^{-1}(\upsilon, p^*) = (w, y). \tag{20}$$

Equations (19) and (1.13b) define the inverse mapping:

$$\mathscr{F}^{-1}(w, y) = (\upsilon, p^*). \tag{21}$$

In accordance with (2) we define

$$G(w, y) = \sum_{k=1}^{m} w_k \tilde{\upsilon}_k(w, y). \tag{22}$$

We shall also define the *indirect production function f* by

$$\tilde{f}(w, y) = f[\tilde{\upsilon}_1(w, y), \tilde{\upsilon}_2(w, y)], \ldots, \tilde{\upsilon}_m(w, y)] \tag{23}$$

which satisfies the identity

$$\tilde{f}(w, y) = y \text{ for all } w, y. \tag{24}$$

Theorem 5. (*Fundamental Envelope Theorem of Production Theory*). The functions G, $\tilde{v}_l$, $\tilde{p}^*$ of (22), (19), (1.13b) are related by

$$\frac{\partial G(w, y)}{\partial w_j} = \tilde{v}_j(w, y) \quad j = 1, 2, \ldots, m \tag{25}$$

and

$$\frac{\partial G(w, y)}{\partial w_j} = \tilde{p}^*(w, y) \tag{26}$$

Proof: Differentiating (15) with respect to w_j, we obtain

$$\frac{\partial G(w, y)}{\partial w_j} = \tilde{v}_j(w, y) + \sum_{k=1}^{m} w_k \frac{\partial \tilde{v}_k(w, y)}{\partial w_j} \tag{27}$$

To prove (25) we must show that the second term on the right of (27) vanishes. Differentiating (23) with respect to w_j and making use of the identity (24) and the chain rule, we obtain upon substitution of (1.13b),

$$0 = \frac{\partial f(w, y)}{\partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial v_k}\bigg|_{v k = \tilde{v} k(w,\, y)} \cdot \frac{\partial \tilde{v}_k(w, y)}{\partial w_j}$$
$$= \frac{1}{\tilde{p}^*(w, y)} \sum_{k=1}^{m} w_k \frac{\partial \tilde{v}_k(w, y)}{\partial w_j}.$$

and (25) follows. Likewise, differentiating (23) with respect to y and using the identity (24) and the chain rule, we have upon making use once again of (1.13b),

$$1 = \frac{\partial \tilde{f}(w, y)}{\partial y} = \sum_{k=1}^{m} \left(\frac{\partial f}{\partial v_k}\right)_{v k = \tilde{v} k(w,\, y)} \cdot \frac{\partial \tilde{v}_k(w, y)}{\partial y}$$
$$= \frac{1}{\tilde{p}^*(w, y)} \sum_{k=1}^{m} w_k \frac{\partial \tilde{v}_k(w, y)}{\partial y}.$$

Thus, from this result and (22),

$$\frac{\partial G(w, y)}{\partial y} = \sum_{k=1}^{m} w_k \frac{\partial \tilde{v}_k(w, y)}{\partial y} = \tilde{p}^*(w, y),$$

establishing (26).

q.e.d.

It may be noted immediately from (22) and (25) that

$$G(w, y) = \sum_{k=1}^{m} w_k \frac{\partial G(w, y)}{\partial w_k},$$

providing the necessary and sufficient condition, by Euler's theorem, that G be homogeneous of degree 1 in w – a result already obtained in §1.1. Using (25) again it follows that $\tilde{v}_j$ is homogeneous of degree zero in w.

Now let us introduce condition (d): the positive homogeneity (of degree 1) of the production function f. Using (22) and (1.13b) we have, by Euler's theorem,

$$G(w, y) = \sum_{k=1}^{m} w_k \tilde{v}_k\,(w, y)$$
$$= \tilde{p}^*(w, y) \sum_{k=1}^{m} \left(\frac{\partial f}{\partial v_k}\right)_{v k = \tilde{v} k(w,\, y)} \cdot \tilde{v}_k(w, y)$$
$$= y \tilde{p}^*(w, y)$$

whence from (6)

$$\tilde{p}^*(w, y) = \frac{G(w, y)}{y} = g(w). \tag{28}$$

Defining

$$b_j(w) = \frac{\partial g(w)}{\partial w_j} \quad (j = 1, 2, \ldots, m) \tag{29}$$

we have from (25), (28), and (29),

$$\tilde{v}_j(w, y) = \frac{\partial G(w, y)}{\partial w_j} = y \frac{\partial g(w)}{\partial w_j} = y b_j(w) \tag{30}$$

hence the optimal factor-product ratios are given by

$$\frac{v_j}{y} = b_j(w). \tag{31}$$

From the differentiability assumption (e) imposed on the function f we can derive a strict quasi-concavity property of the function g. For suppose $w^0 > 0,\ w^1 > 0,\ and\ w^0 \neq \lambda w^1$; then from ($b_2$) and (e), we have $b(w^0) \neq b(w^1)$, where

$$b(w) = [b_1(w), b_2(w), \ldots, b_a(w)]. \tag{32}$$

Now by definition of g [see (5)]

$$\begin{aligned} g(w^0) &\leqq w^0 \cdot \upsilon \text{ for all } \upsilon \in A(1) \\ g(w^1) &\geqq w^1 \cdot \upsilon \text{ for all } \upsilon \in A(1) \end{aligned} \tag{33}$$

and moreover

$$\begin{aligned} g(w^0) &= w^0 \cdot \upsilon \text{ if and only if } \upsilon \in b(w^0) \\ g(w^1) &\geqq w^1 \cdot \upsilon \text{ if and only if } \upsilon \in b(w^0) \end{aligned} \tag{34}$$

Furthermore, $b(w^0) \neq b(w^1)$, so strict inequality must hold in one of the inequalities (33); thus if $0 < \theta < 1$,

$$(1-\theta)g(w^0) + \theta g(w^1) < \left[(1-\theta)w^0 + \theta w^1\right] \cdot v \text{ for all } v \in A(1)$$

and therefore in particular

$$\begin{aligned} &(1-\theta)g(w^0) + \theta g(w^1) \\ &< \min_v \left\{ \left[(1-\theta)w^0 + \theta w^1\right] \cdot v : v \in A(1) \right\} \\ &= g[(1-\theta)w^0 + \theta w^1]. \end{aligned}$$

So we have

(b^*) if $w^0 > 0,\ w^1 > 0$, and $w^0 \neq \lambda w^1$, and if $0 < \theta < 1$, then $g[(1-\theta)w^0 + \theta w^1] > (1-\theta)g(w^0) + \theta g(w^1)$.

If is not hard to see that a corresponding property (b_3) holds for f as well. Failure of $\left(b_3^*\right)$ when f is not differentiable, allowing $b(w^0) = b(w^1)$ for $w^0 \neq \lambda w^1$, is illustrated in Fig. 4.

In general, a flat segment on a production isoquant goes over into a kink on the dual cost isoquant, and vice versa. There is another still more subtle relationship, illustrated by the following function found in Katzner (1970, p. 54):

$$f(\upsilon_1, \upsilon_2) = \left(\upsilon_1^3 \upsilon_2 + \upsilon_1 \upsilon_2^3\right)^{1/4}$$

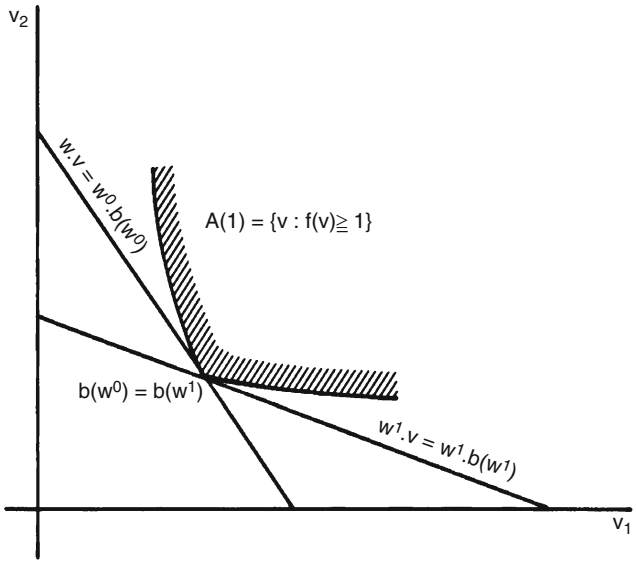
Its dual minimum-unit-cost function is found to be

$$g(w_1, w_2) = 2^{-1/4}\left[(w_1 + w_2)^{4/3} - (w_1\, w_2)^{4/3}\right]^{4/3}.$$

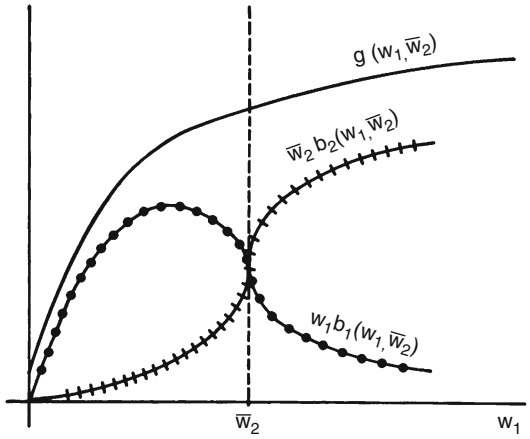
The isoquants of f are extremely flat at $\upsilon_1 = \upsilon_2$, and as a result g is once but not twice differentiable at $w_1 = w_2$. A graph of

$$g(w_1, w_2) = w_1 b_1(w_1,\ w_2) + w_2 b_2(w_1\ w_2)$$

for $w_2 = \overline{w}_2$ is shown in Fig. 5. At $w_1 = \overline{w}_2$, $\overline{w}_2$ $b_2(w_1,\ \overline{w}_2)$ has a slope of $+\infty$ and $w_1 b_1(w_1,\ \overline{w}_2)$ has a slope of $-\infty$, yet their sum is differentiable.



International Trade, Fig. 4



International Trade, Fig. 5

When the bordered Hessian of the production function f is invertible, its inverse is the bordered Hessian of the cost function g; in the above example, it is not invertible at $\upsilon_1 = \upsilon_2$.

A useful illustration of the duality of cost and production functions is given by the case of CES (constant-elasticity-of-substitution) production functions (cf. Arrow et al. 1961; Uzawa 1962):

$$f(\upsilon) = \left[\sum_{i=1}^{m} \alpha_i \upsilon_i^{1-1/\sigma}\right]^{\sigma/(\sigma-1)}$$

The corresponding cost functions have the form

$$f(w) = \left[\sum_{i=1}^{m} \alpha_i^{\sigma} w_i^{1-\sigma}\right]^{1/(1-\sigma)}$$

whose elasticity of substitution is $\sigma^* = 1/\sigma$.

The Production-Possibility Set

Suppose a country to be capable of producing n commodities with the aid of m primary factors of production. Denoting the output of commodity j by y_j, and the input of factor i into the production of commodity j by υ_{ij}, the production function may be written

$$\begin{aligned} y_j &= f_i(\upsilon_{1j}, \upsilon_{2j}, \ldots, v_{mj}) \\ &= f_j(\upsilon_j) \qquad (j = 1, 2, \ldots, n), \end{aligned} \tag{35}$$

where

$$\upsilon_{.j} = (\upsilon_{1j}, \upsilon_{2j}, \ldots, \upsilon_{mj}). \tag{36}$$

It will be assumed that f_j is:

(a) *Continuous*; i.e.,

$$\lim_{\upsilon_j \to \upsilon_j^0} f_j(\upsilon_j) = f_j\left(\upsilon_j^0\right);$$

(b) *Weakly monotone*; i.e., if $\upsilon^1{}_{.j} \geqq \upsilon^2{}_{.j}$ (meaning that $\upsilon^1{}_{ij} \geqq \upsilon^2{}_{ij}$ for $i = 1, 2, \ldots, m$) then $f_i(\upsilon^1{}_{.j}) \geqq f_j(\upsilon^2{}_{.j})$, and if $\upsilon^1{}_{.j} > \upsilon^2{}_{.j}$ (i.e., $\upsilon^1{}_{ij} > \upsilon^2{}_{ij}$ for $i = 1, 2, \ldots, m$) then $f_i(\upsilon^1{}_{.j}) > f_j(\upsilon^2{}_{.j})$;

(c) *Concave*; i.e., if $\upsilon^0{}_{.j}$ and $\upsilon^1{}_{.j}$ are any two vectors of primary inputs into the production of commodity j, then for any t in the interval $0 < t < 1$, (Fig. 5)

$$f_j\left[(1-t)\upsilon_j^0 + t\upsilon_j^1 \geqq (1-t) f_j\left(\upsilon_j^0\right) + t f_j\left(\upsilon_j^1\right)\right. \tag{37}$$

(d) *Positively homogeneous of degree* 1; i.e., for any $\lambda > 0$,

$$f_j(\lambda \upsilon_{.j}) = \lambda f_j(\upsilon_{.j}). \tag{38}$$

It will be convenient to introduce the $m \times n$ allocation matrix

$$V = \begin{bmatrix} \upsilon_{11} & \upsilon_{12} & \cdots & \upsilon_{1n} \\ \upsilon_{21} & \upsilon_{22} & \cdots & \upsilon_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \upsilon_{m1} & \upsilon_{m2} & \cdots & \upsilon_{mn} \end{bmatrix} \tag{39}$$

The element υ_{ij} is the input of factor i into the production of commodity j. The j the column of V will be donoted υ_j; according to this notation, υ_j is the transpose of $\upsilon_{.j}$, denoted $\upsilon_j = \mathrm{v}'_{.j}$.

Let l_i denote the country's total endowment of factor i. Then for each i the following resource constraint holds:

$$\sum_{j=1}^{m} v_{ij} \leqq l_j \quad (i = 1, 2, \ldots m). \tag{40}$$

Using (39) this can be written in matrix notation as

$$\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \leqq \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_m \end{bmatrix}, \tag{41}$$

or simply

$$V_i \leqq l, \tag{42}$$

where l is the column vector of n ones and $l = (l_1, l_2, \ldots, l_m)'$ is the column vector of factor endowments.

In the absence of any additional restrictions, condition (40) expresses the *perfectly mobility* of factors among industries.

The country's *production–possibility set* is the set of all possible output combinations $y = (y_1, y_2, \ldots, y_n)$ that can be produced with the production functions (35) under the resource constraints (40). Formally, it may be denoted

$$\begin{aligned} \mathscr{Y}(l) = y : \text{there exist allocations } v_{ij} \geqq 0 \textit{ such that} \\ y_j = f_j(v_j)(j = 1, 2, \ldots n) \quad \text{and} \\ \sum_{j=1}^{n} v_{ij} \leqq l_j \quad (i = 1, 2, \ldots m). \end{aligned} \tag{43}$$

For notational convenience we may define the function $f(V)$ as the vector-valued function

$$f(V) = \left(f_1\left(v_1'\right), f_2\left(v_2'\right), \ldots, f_n\left(v_n'\right)\right)' \tag{44}$$

and write (43) in the more compact form

$$y(l) = \{y : (\exists V \geqq 0) y = f(V) \& V_l \geqq l\}. \tag{45}$$

Note that with this notation, condition (37) can be written (for $t = t_j$) in the form

$$\begin{aligned} f\left(V^0(I - T) + V^1 T\right) \geqq (I - T) f\left(V^0\right) \\ + T f\left(V^1\right) \end{aligned} \tag{46}$$

where $T = \text{diag}(t_1, t_2, \ldots, t_n)$ is an $n \times n$ diagonal matrix with $0 < t_j < 1$. Likewise, (38) may be written (for $\lambda = \lambda_j$) in the form

$$f(V\Lambda) = \Lambda f(V), \tag{47}$$

where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is an $n \times n$ diagonal matrix with $\lambda_j > 0$.

Theorem 6. If assumptions (a), (b), and (c) hold, the production-possibility set $\mathscr{Y}(l)$ is convex.

Proof: Let y^0, y^1 both belong to $\mathscr{Y}(l)$; we are to show that for any t in the interval $0 < t < 1$, the output combination $y^t = (l - t)\, y^0 + ty^1$ also belongs to $\mathscr{Y}(l)$ (see Fig. 6).

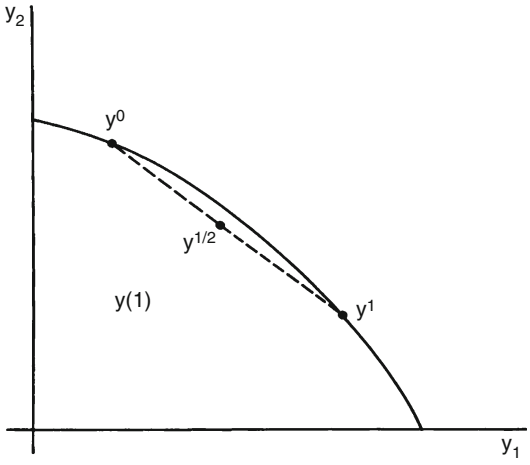
Since $y^0, y^{ty} \in \mathscr{Y}(l)$, this means that there exist two allocation matrices V^0, V^1 each satisfying (42), such that $y^0 = f(V^0)$ and $y^1 = f(V^1)$. Denote $V^t = (1 - t)V^0 + tV^0$. Then from (42),

$$V^t l = (1 - t)V^0 t + tV^1 l \leqq (1 - t)l + tl = l, \tag{48}$$

so V^t is a feasible allocation, and by concavity,

$$f(V^t) \geqq (1 - t) f\left(V^0\right) + tf\left(V^1\right) = y^t, \tag{49}$$

i.e., for each $j = 1, 2, \ldots, n$, denoting $v^t{}_{.j} = (1 - t)^{V}{}_{.j} + tv^1{}_{.j}$,



International Trade, Fig. 6

$$f_j\left(v^t{}_{\cdot j}\right) \geqq (1-t) f_j\left(v^0{}_{\cdot j}\right) + t f_j\left(v^1{}_{\cdot j}\right) = y^t{}_j. \quad (50)$$

By continuity and monotonicity of f_j, there exist $\lambda^t{}_j \leqq 1$ such that

$$f_j\left(\lambda^t_j v^t_j\right) = y^t_j \quad (j = 1, 2, \ldots, n). \quad (51)$$

(In particular, (51) follows if the stronger homogeneity condition (d) holds, by taking $\lambda^t{}_j = y^t_j / f_j\left(v^t_j\right)$ if $y^t{}_t > 0$, and 0 otherwise.) Equivalently,

$$f(V^t \Lambda^t) = y^t. \quad (52)$$

It remains only to verify that the matrix $V^t\Lambda$ of allocations $\lambda^t_j\, v^t_{ij}$ satisfies the constraint (42).

This is immediate from the fact that $0 \leqq \lambda^t_j \leqq 1$, whence from (48),

$$V^t \Lambda^t_l = V^t \lambda^t \leqq V^t_l \leqq l. \quad (53)$$

q.e.d.

Note that homogeneity of production functions is not needed for the above result (Fig. 7).

The National-Product Function

Let $p = (p_1, p_2, \ldots, p_n)'$ denote a vector of prices. The *national-product function* (cf. Samuelson 1953; Chipman 1972, 1974) is defined as the function

$$\Pi(p, l) = \max_{y \in \mathscr{Y}(l)} p \cdot y. \quad (54)$$

[See also Dixit and Norman (1980), who use the terminology 'revenue function'.]

For any fixed p, this has all the properties of a production function, but with some special peculiar features. These are illustrated in Fig. 7 to be explained shortly.

For each commodity, $j = 1, 2, \ldots, n$, define the uppercontour set

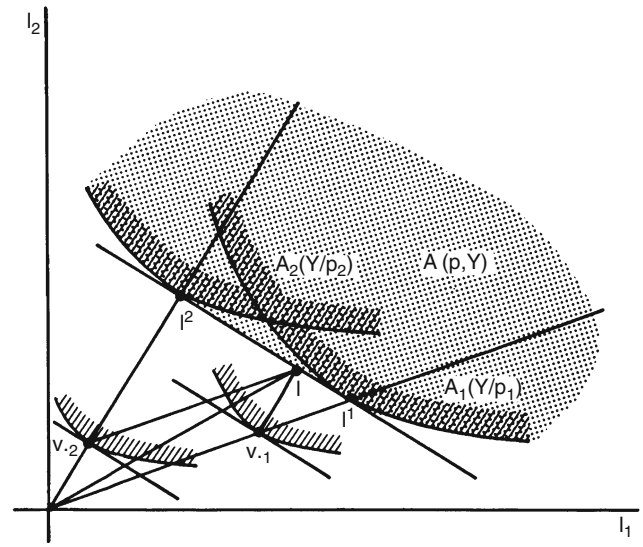
$$A_j\left(y_j\right) = \left\{l^j = \left\{l^j_1 = \left(l^j_1, l^j_2, \ldots l^j_m\right) : f_j\left(l^j\right) \geqq y_j\right\}\right. . \quad (55)$$

Then in particular,

$$A_j(Y/p_j) = \left\{l^j : p_j f_j\left(l^j\right) \geqq Y\right\} \quad (56)$$

is the set of factor-input combinations that will yield, at the given price p_j an amount of commodity j worth at least Y. Throughout this section it will be assumed that each f_j satisfies properties (a)–(d) of the preceding section.

Let us now introduce a stronger monotonicity condition that refers to the entire vector-valued function (63). It may be stated as follows: f is



International Trade, Fig. 7

(e) *Strictly monotone*, i.e., for each $V = [\upsilon_{ij}]$ and each $i = 1, 2, \ldots, m$, there is a $j = 1, 2, \ldots, n$ such that $\delta > 0$ implies

$$f_j(\upsilon_{1j}, \upsilon_{2j}, \ldots, \upsilon_{ij} + \delta, \ldots, \upsilon_{mj}) > f_j(\upsilon_{1j}, \upsilon_{2j}, \ldots, \upsilon_{ij}, \ldots, \upsilon_{mj}). \tag{57}$$

In words, if there is an increase in the amount of any one of the m endowments, it is possible to find an industry where this additional input will lead to increased output.

For any family of sets $S_1, S_2, \ldots, S_n$, each a subset of m-dimensional Euclidean space E^m, the *arithmetic mean* of this family (which is, for convex S_j, also the *convex hull* of $\cup^n{}_{j=1} S_j$) is defined and denoted

$$\mathop{\mathrm{M}}_{j=1}^{n} S_j = \left\{ s \in E^m : \left(\exists s^j \in S_j,\ \lambda_j \geqq 0, j = 1, 2, \ldots, n\right) \sum_{j=1}^{n} \lambda_j = 1 \text{ and } s = \sum_{j=1}^{n} \lambda_j s^j \right\} \tag{58}$$

Analogously to (55) we define the upper-contour set of the national-product function by

$$A(p, Y) = \left\{ l \in E_+^m : \prod(p, l) \geqq Y \right\}. \tag{59}$$

The following theorem characterizes the isoquants of the function $\Pi(p, \cdot)$ (see Fig. 7).

Theorem 7. Let all prices p_j, be positive, $j = 1,2,\ldots, n$, and let f satisfy conditions (a)–(d) of section "The Production-Possibility Set", as well as the strict montonicity condition (e). Then

$$A(p, Y) = \mathop{\mathrm{M}}_{j=1}^{n} A_j(Y/p_j), \tag{60}$$

i.e., the upper-contour set consisting of all factor combinations l that give rise to a national product of at least Y, is the arithmetic mean of the n upper-contour sets consisting, for each commodity j, of all factor combinations l^j that, when allocated entirely to industry j, give rise to a national product of at least Y.

Proof (a) let us first prove that

$$\mathop{\mathrm{M}}_{j=1}^{n} A_j(Y/p_j) \subseteq A(p, Y) \tag{61}$$

Let

$$l \in \mathop{\mathrm{M}}_{j=1}^{n} A_j(Y/p_j).$$

Then, by definition (58), there exist $l^j \in A(Y/p_j)$ and $\lambda_j \geqq 0$ such that

$$\sum_{j=1}^{n} \lambda_j = 1 \quad \text{and} \quad \sum_{j=1}^{n} \lambda_j l^j = l.$$

By definition (56), each l^j satisfies $p_j f_j(l^j) \geqq Y$, hence from the definition (54) of Π and the homogeneity of degree 1 of each f_j, we have

$$\Pi(p, l) \geqq \sum_{j=1}^{n} p_j f_j(\lambda_j l^j) = \sum_{j=1}^{n} p_j f_j(l^i) \geqq y \sum_{j=1}^{n} \lambda_j = Y.$$

From definition (59) it follows that $l \in A(p, Y)$, and (61) follows.

(b) We now show that

$$A(p, Y) \subseteq \mathop{\mathrm{M}}_{j=1}^{n} A_j(Y/p_j). \tag{62}$$

Let $l \in A(p, Y)$; then by definitions (59), (54) and (43), there exist allocations $\upsilon_j \in E^{m+}$ such that

$$\sum_{j=1}^{n} \upsilon_j \leqq l \quad \text{and} \quad \sum_{j=1}^{n} p_j f_j(\upsilon_j) = \Pi(p, l) \geqq Y. \tag{63}$$

By the strict monotonicity of f, the first inequality of (63) must be an equality; for, if for some $i = i'$ we have

$$\sum_{j=1}^{n} v_{i'j} < l_{i'}$$

then for some

$$j = j' \quad \text{and} \quad 0 < \delta \leqq l_{i'} - \sum_{j=1}^{n} v_{i'j}$$

the inequality (57) is satisfied, violating the definition (54) of $\Pi(p, l)$. Now define

$$\begin{aligned} \lambda_j &= p_j f_j(v_{\cdot j}) / \Pi(p, l), \\ l^j &= u_{\cdot j} / \lambda_j \quad (j = 1, 2, \ldots, n). \end{aligned} \tag{64}$$

Then

$$\sum_{j=1}^{n} \lambda_j l^j = 1 \quad \text{where} \quad \sum_{j=1}^{n} \lambda_j = 1 \tag{65}$$

By homogeneity we have

$$p_j f_j(l^j) = p_j f_j(v_{\cdot j}) / \lambda_j = \Pi(p, l) \geqq Y,$$

hence $l^j \in A_j(Y/p_j)$ from (56). Together with (63) this implies that (62) holds.

q.e.d.

Since for each fixed p the national-product function $\Pi(p, \ )$ has the properties of a production function (i.e. it is continuous, concave, monotone, and positively homogeneous of degree 1), we may associate with it a corresponding minimum-unit cost function $\Gamma(p, \ )$ defined by

$$\Gamma(p, w) = \min_{l} \left\{ w \cdot l : \prod(p, l) \geqq \right\}. \tag{66}$$

This will be called the *national-cost function*. Letting $g_j(w) = min_{v_{\cdot j}} \{w. v_{\cdot j} : f_j(v_{\cdot j}) \geqq 1\}$ denote the minimum-unit cost function dual to the production function $f_j(v_{\cdot j})\}$, we may define the upper-contour sets

$$A_j^*(p_j) = \left\{ w : g_j(w) \geqq p_j \right\} \tag{67}$$

And

$$A^*(p) = \{ w : \Gamma(w) \geqq 1 \} \tag{68}$$

The boundary of the intersection of all the sets (67) for $j = 1, 2, \ldots, n$ is known as the 'factor-rental frontier' (or 'factor-price frontier'-cf. Woodland 1982, pp. 49–52). The following theorem shows that it is also the contour of the corresponding national-cost function. Its shape will be similar to that depicted in Fig. 4.

Theorem 8. Let the prices p_j be positive, $j = 1, 2, \ldots, n$ and let f satisfy conditions (a) to (e) of section 1.2. Then

$$A*(p) = \bigcap_{j=1}^{n} A_j^*(p_j). \tag{69}$$

Proof Let $w \in A^*(p)$; then $\Gamma(p, w) \geqq 1$, i.e., $w \, l \geqq 1$, for all $l \in A(p, l)$. Choose such an l and let V be the optimal resource-allocation matrix: then

$$\Pi(p, l) = \sum_{j=1}^{n} p_j f_j(v_j) \geqq 1. \tag{70}$$

Defining λ_j and l^j as in (64), this gives (by homogeneity)

$$\sum_{j=1}^{n} p_j f_j(\lambda_j l^j) = \sum_{j=1}^{n} \lambda_j p_j f_j(l^j) \geqq 1, \tag{71}$$

and since

$$\lambda_j > 0 \text{ and } \sum_{j=1}^{n} \lambda_j = 1$$

this implies $p_j f_j(l^j) \geqq 1$, i.e., $l^j \in A_j(1/p_j)$, for each j. Now by hypothesis, (70) implies $w \geqq l \geqq 1$ hence

$$\sum_{j=1}^{n} \lambda_j w \cdot l^j \geqq 1, \tag{72}$$

and by the same reasoning as above this implies $w. l^j \geqq 1$ for all j, i.e.,

$$g_j(w)/p_j = \min_{l^j} \left\{ w \cdot l^j : l^j \in A_j(1/p_j) \right\} \geqq 1 \tag{73}$$

or $g_j(w) \geqq p_j$. From the definition (67) this shows that $w \in A_j^*(p_j)$ *for* $j = 1, 2, \ldots, n$.

Conversely, let $w \in \cap_{j=1}^{n} A_j^*(p_j)$; then $g_j(w) \geqq p_j$ for $j = 1, 2, \ldots, n$. From the definition of g_j, this implies $w\, l^j \geqq 1$ for all $l^j \in A_j(l/p_j), j = 1, 2, \ldots, n$. Choosing $l^j \in A_j\,(1/p_j)$ such that

$$\sum_{j=1}^{n} \lambda_j l^j = l, \Pi(p,l) \geqq \sum_{j=1}^{n} p_j f_j(\lambda_j l^j) = \sum_{j=1}^{n} \lambda_j p_j f_j(l^i) \geqq \sum_{j=1}^{n} \lambda_j = 1, \tag{74}$$

Hence

$$w \cdot l = w \cdot \sum_{j=1}^{n} \lambda_j l^j = \sum_{j=1}^{n} \lambda_j w \cdot l^j \geqq 1, \tag{75}$$

From the definition (66) this implies $\Gamma(p, w) \geqq 1$, and thus by (68) it follows that $w \in A^*(p)$. q.e.d.

Let us introduce a further assumption, that each f_j is

(f) *Differentiable.*

Then from Theorem 7 it follows that $\Pi(p, .)$ is differentiable. Its partial derivative with respect to l_i *is defined as the Stolper–Samuelson Function*

$$\widehat{w}_i(p,l) \equiv \frac{\partial}{\partial l_i} \Pi(p,l) (i = 1, 2, \ldots, m), \tag{76}$$

and the corresponding vector-valued function $\widehat{w}_i(p,l) \equiv \partial \prod(p,l)/\partial l$ is called the *Stolper-Samuelson mapping*. The values of this function are the shadow or implicit factor rentals of the respective factors.

Setting up the Lagrangean function

$$L(V, w; p, l) = \sum_{j=1}^{n} p_j f_j(\upsilon_j) - \sum_{i=1}^{m} w_i \left(\sum_{j=1}^{n} \upsilon_{ij} - l_i \right) \tag{77}$$

corresponding to the definition of the national-product function, we obtain the Kuhn-Tucker conditions

$$\frac{\partial L}{\partial \upsilon i_j} = p_j \frac{\partial f_i}{\partial \upsilon_{ij}} - w_i \leqq 0; \quad \left(p_j \frac{\partial f_i}{\partial \upsilon_{ij}} - w_i \right) \upsilon_{ij} = 0; \tag{78}$$

$$\frac{\partial L}{\partial w_i} = l_i - \sum_{j=1}^{n} \upsilon_{ij} \geqq 0, \quad \left(l_i - \sum_{j=1}^{n} \upsilon_{ij} \right) w_i = 0. \tag{79}$$

It will be observed that conditions (78) constitute, for each $j = 1,2,\ldots,n$, precisely the Kuhn–Tucker conditions for costminimization in industry j, where w_i, is the ith factor rental. The rentals defined by the Stolper–Samuelson mapping are therefore the market rentals that will obtain in competitive equilibrium.

Let us now explore the consequences of assuming that the function Π is differentiable with respect to p as well as l. Given p^0, l^0, let y^0 maximize $p^0\, y^0$. Define the function

$$H(p, l^0) = \Pi(p, l^0) - p \cdot y^0.$$

Then $H(p^0, l^0) = 0$ and $H(p, l^0) \geqq 0$ for $p \neq p^0$ (by the definition of Π), hence H reaches a minimum with respect to p at $p = p^0$. Since differentiability of Π implies differentiability of H, we have

$$\frac{\partial H(p^0, l^0)}{\partial p_j} = \frac{\partial \Pi(p^0, l^0)}{\partial p_j} - y_j^0 = 0.$$

This shows that y^0 is the *unique* y which maximizes $p^0.y$ subject to $y \in \mathscr{Y}\ (l^0)$. This is equivalent to saying that the *production-possibility frontier* $\widehat{\mathscr{Y}}$ (l) – i.e., the set of all $y \in \mathscr{Y}\ (l^0)$ which maximize $p.y$ for some $p > 0$ – is *strictly concave to the origin*. The apparently innocuous assumption that Π is differentiable with respect to p has thus led to an important substantive conclusion.

When Π is differentiable with respect to p, the function

$$\widehat{y}_j(p,l) = \frac{\partial}{\partial p_j} \Pi(p,l) \quad (j = 1, 2, \ldots, n) \tag{80}$$

is called the *Rybczynski function* for commodity j. The corresponding vector-valued function $\hat{y}(p, l)$ is called the *Rybczynski mapping*.

In general, we may define the *Rybczynski correspondence* by

$$\widehat{y}(p,l) = \left\{ y \in y(l) : p \cdot y = \prod(p,l) \right\}. \quad (81)$$

The above result shows that if Π is differentiable with respect to p, this correspondence is a singleton-valued mapping. We shall now obtain a necessary and sufficient condition for this single-valuedness, i.e., for the strict concavity to the origin of y (1).

Let the factor–output coefficients be denoted

$$b_{ij}(w) = \frac{\partial g_j(w)}{\partial w_i} (i = l,2,\ldots,m;\ \ j = 1,2,\ldots,n) \quad (82)$$

where g_j is the minimum-unit-cost function dual to the production function f_j. The following result was obtained by Khang (1971) and Chipman (1972).

Theorem 9. Let p^0, l^0 be such that there exists a $y^0 > 0$ which maximizes $p^0 \cdot y$ subject to $y \in \mathscr{Y}$ (l^0), and let $w^0 = \hat{w}\ (p^0, l^0) = \partial \Pi(\mathrm{p}^0, \mathrm{l}^0)/\partial l$. Let f satisfy the strict monotonicity condition (e).

Then in order that y^0 should be the unique maximizer of $p^0\ y$ subject to $y \in \mathscr{Y}\ (l^0)$, it is necessary and sufficient that the n columns of the factor-output matrix

$$B(w^0) = \begin{bmatrix} b_{11}(w^0) & b_{12}(w^0) & \cdots & b_{1n}(w^0) \\ b_{21}(w^0) & b_{22}(w^0) & \cdots & b_{2n}(w^0) \\ & & & \\ b_{m1}(w^0) & b_{m2}(w^0) & \cdots & b_{mn}(w^0) \end{bmatrix}$$

be linearly independent.

Proof: For convenience, denote $B^0 = B(w^0)$. Then from strict monotonicity of f we have

$$B^0 y^0 = l^0. \quad (83)$$

First we show that if rank $B^0 < n$ then y^0 is not unique. Since rank $B^0 < n$ there exists a vector $z^0 \neq 0$ such that

$$B^0 z^0 = 0. \quad (84)$$

Choose $\varepsilon^0 > 0$ such that

$$y^0 \pm \varepsilon^0 z^0 > 0;$$

then $y^0 \pm \varepsilon\ z^0 > 0$ for $0 < \varepsilon < \varepsilon^0$. From (83) and (84) we have $B^0(y^0 \pm \varepsilon^0 z^0) = l^0$ whence $y^0 \pm \varepsilon^0 z^0 \in \mathscr{Y}\ (l^0)$. Since y^0 maximizes $p^0 \cdot y$ over $\mathscr{Y}\ (l^0)$,

$$p^0 \cdot y^0 \geqq p^0 \cdot \left(y^0 \pm \varepsilon^0 z^0\right),$$

i.e., $0 \geqq \varepsilon^0 p^0 \cdot z^0 \geqq 0$. This implies $p^0 \cdot z^0 = 0$, hence $p^0 \cdot (y^0 \pm \varepsilon z^0) = p^0 y^0$ for $0 < \varepsilon < \varepsilon^0$, i.e.,

$$y^0 \pm \varepsilon z^0 \in y\left(l^0\right) \text{ for } 0 < \varepsilon < \varepsilon^0.$$

This shows that y^0 is not unique.

Conversely we show that if y^0 is not unique then rank $B^0 < n$. Suppose $y^0, y^1 > 0$ both maximize $p^0 \cdot y$ subject to $y \in = (l^0)$, where $y^1 \neq y^0$. Then $B^0 y^0 = B^0 y^1 = l$, hence $B^0(y^0 - y^1) = 0$; since $y^0 - y^1 \neq 0$, this implies that rank $B^0 < n$.

q.e.d.

From this result it follows that a necessary condition for the production-possibility frontier to be strictly concave to the origin is that $m \geqq n$. If $m < n$, it is ruled surface. However, the condition $m \geqq n$ is certainly not sufficient; one example is the case $m = n = 2$ when two isoquants for a dollar's worth of output are mutually tangent at a point along the endowment ray (cf. Lerner 1933, p. 13). For further discussion of these points see Kemp et al. (1978), and for an interesting characterization, see Inoue (1986) and Inoue and Wegge (1986).

To gain an intuitive understanding of the meaning of the differentiability of Π (.,l), let us assume that the f_j are differentiable and that the functions $\widehat{v}_{ij}\ (p, l)$, obtained with the $\hat{w}_i\ (p, 1)$ by solving the above constrained-maximum problem, are also single-valued and differentiable. Then from

$$\Pi(p, l) = \sum_{j=1}^{n} p_j f_j\left(\upsilon_j\right) \quad (85)$$

we have

$$\frac{\partial \Pi}{\partial p_k} = y_k + \sum_{j=1}^{n}\sum_{i=1}^{m}\left[p_j \frac{\partial f_j}{\partial v_{ij}} - w_i\right]\frac{\partial \hat{v}_{ij}}{\partial p_k} + \sum_{i-1}^{m} w_i \sum_{j=1}^{n}\frac{\partial \hat{v}_{ij}}{\partial p_k}. \tag{86}$$

If $w_i > 0$ then

$$\sum_{j=1}^{n} \hat{v}_{ij}(p, l) = l_i$$

$$\sum_{j=1}^{n} \partial \hat{v}_{ij}/\partial p_k = 0,$$

hence the last term of (86) must vanish. If $v_{ij} > 0$ then the bracketed term in (86) vanishes (by the Kuhn-Tucker conditions). If the bracketed term is negative then $\hat{v}_{ij}$ by the Kuhn-Tucker conditions, and thus $\partial \hat{v}_{ij}/\partial p_k = 0$. In either case, the second term on the right in (86) vanishes. The trouble occurs in the intermediate case in which factor i is on the verge of being employed in industry j, hence $p_j \partial f_j/\partial v_{ij} - w_i = 0$ and $v_{ij} = 0$; it is precisely in this case that $\hat{v}_{ij}(\cdot, l)$ will not be differentiable at that point. Formula (80) therefore fails at switching points where factors are on the verge of being employed in particular industries; a small price change in one direction will lead to their continued unemployment, but in the other direction to their being employed. Thus, $\Pi(\cdot,l)$ is non-differentiable at such switching points. Likewise, it is non-differentiable when the conditions of Theorem 9 fail, in which case a small price change may lead to a country's switching from specialization in one commodity to specialization in another. All this would become clearer if the theory were to be recast in terms of subdifferentials (cf. Rockafellar 1970).

Since $\Pi(p, \cdot)$ has the properties of a production function, from Theorem 7, it is concave; and since, as was seen above, $H(P,l^0) = \Pi(p, l^0) - p \cdot y^0$ is a minimum at $p = p^0$, where $\Pi(p^0, l^0) = p^0 \cdot y^0$, $H(\cdot, l)$ is convex, hence $\Pi(\cdot,l)$ is convex. That is, $\Pi(p, l)$ is convex in p and concave in l.

If it is twice continuously differentiable then Samuelson's (1953) 'reciprocity theorem' holds:

$$\frac{\delta \hat{y}_j}{\delta l_j} = \frac{\partial \prod}{\partial p_j \partial l_i} = \frac{\partial^2 \prod}{\partial l_i \partial p_j} = \frac{\partial \hat{w}_i}{\partial p_j}. \tag{87}$$

The Stolper–Samuelson And Rybczynski Mappings

When a country diversifies its production, by which we shall mean that it produces all n consumable commodities, as long as it is not on the verge of specializing, its factor-endowment vector will lie in the interior of a diversification cone – the convex cone whose extreme rays pass through the factor-input vectors in the n industries which minimize costs at the given factor rentals (cf. McKenzie 1955; Chipman 1966). As is clear from Fig. 7, the factor rentals will remain unchanged as the factor endowment vector varies within the interior of this cone; i.e. the function $\hat{w}(p, l)$ is independent of l for endowments l in this cone. Now if all n commodities are to be producted, costs cannot exceed prices; and competitive equilibrium requires that prices not exceed costs. Hence, from the homogeneity of degree 1 of the minimum-unit-cost functions, and by Theorem 5, we have

$$p_j = g_j(w) = \sum_{i=l}^{m} \frac{\partial g_j(w)}{\partial w_i} w_i = \sum_{i=l}^{m} b_{ij}(w) w_i, \tag{88}$$

or in matrix notation, where w and p denote column vectors of m factor rentals and n commodity prices respectively,

and thus

$$p = g(w) = B(w)' w \tag{89}$$

$$g[\hat{w}(p, l)] = p, \tag{90}$$

i.e., $\hat{w}(\cdot,l)$ is a local inverse of the mapping g. Since the Jacobian matrix of g,$B(w)'$, must have rank m if the diversification cone has a non-empty

interior (hence $n \geqq m$), the range of g is an m-dimensional manifold, hence (89) implies that the vector p of world prices cannot be varied arbitrarily (if the country is to continue to diversify) unless $n = m$.

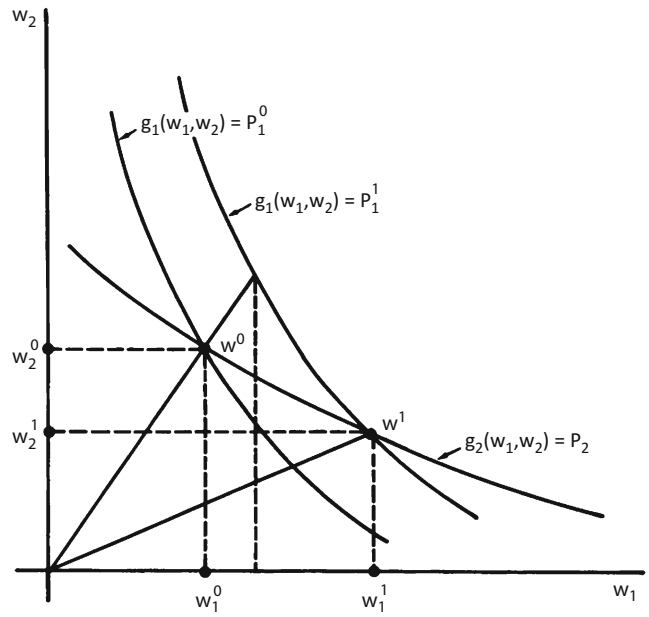
Even when $n = m$, the mapping g is in general not globally univalent. Gale and Nikaido (1965) obtained strong sufficient conditions for global univalence, namely that the principal minors of $B(w)$ be positive (this condition can be slightly weakened). Inada (1971) obtained some alternative conditions. In a controversy with Pearce (1967), McKenzie (1967) showed that it did not suffice to assume that $B(w)$ had a non-vanishing determinant for all positive w. The condition that $|B(w)| \neq 0$ for some $w = w^0$ is of course sufficient for local invertibility of g, but this inverse mapping depends on l. If two countries with identical technologies have their endowment vectors l in the same diversification cone, their factor rentals will be equalized even if g is not globally univalent. Nikaido (1972) showed that a modification of conditions originally suggested by Samuelson (1953) is sufficient for global univalence of g (Fig. 8).

Of particular interest is the nature of the Stolper-Samuelson mapping in regions where it is locally independent of l, i.e., the nature of the local inverses of g. For the reasons given above, discussion of this is effectively limited to the case $n = m$. Defining the diagonal matrices $W = \text{diag } w$ and $P = \text{diag } p$, and the matrix $\mathrm{B} = WBP^{-1}$, by dividing (88) through by p_j, one sees that B is column-stochastic (i.e., has unit column sums in addition to having non-negative elements); denoting its elements by $\beta_{ij} = w_i b_{ij}/p_j = \partial \log g_j/\partial \log w_i$, these satisfy

$$\sum_{i=l}^{m} \beta_{ij} = 1.$$

Denoting the elements of B^{-1} by b^{ij} and those of $B^{-1} = PB^{-1}W^{-1}$ by $\beta^{ij} = p_i b^{ij}/w_j$, these are equal to $\beta^{ij} = \partial \log \hat{w}_j/\partial \log p_i$. Denoting by l_m the column vector of m 1 s, from $l'_m \mathrm{B} = l'_m$ we have $l'_m \mathrm{B}^{-1} = l'_m \mathrm{B}\mathrm{B}^{-1} = l'_m$, hence B^{-1} also has unit column sums (cf. Chipman 1969, p. 402).

In the case $m = n = 2$, if we follow the convention of numbering commodities and factors in such a way that, at the initial equilibrium, $|B(w)| > 0$, i.e.,



International Trade, Fig. 8

$$\begin{vmatrix} b_{11}(w)\, b_{12}(w) \\ b_{21}(w)\, b_{22}(w) \end{vmatrix} = b_{11}(w) b_{12}(w) \left[\frac{b_{22}(w)}{b_{12}(w)} - \frac{b_{21}(w)}{b_{11}(w)} \right] > 0 \tag{91}$$

(which means that industry 2 uses a higher ratio of factor 2 to factor 1 than industry 1), then B^{-1}, which has non-positive diagonal elements and unit column sums, must have diagonal elements greater than or equal to unity. If B has its elements all positive, then the off-diagonal elements of B^{-1} are negative and the diagonal elements greater than unity. This, in substance, is the Stolper-Samuelson (1941) theorem. In words, for some association of commodities and factors, a rise in

one commodity price will lead to a more than proportionate rise in the corresponding factor rental (Fig. 9).

A simple proof is illustrated in Fig. 8, in the space of factor rentals. A rise in p_1 is shown by an upward shift in the isoquant $g_1(w_1, w_2) = p_1$ and a new intersection point with the isoquant $g_2(w_i, w_2) = p_2$ with lower w_2 and the rise in w_1 proportionately higher than that of p_1 (as long as the elasticities of substitution of the cost functions are positive, i.e., as long as the elasticities of substitution of the production functions are finite).

The Stolper-Samuelson theorem clearly does not generalize to higher dimensions. Either much stronger assumptions or much weaker conclusions are required. See Chipman (1969), Kuhn (1968), Inada (1971), Uekawa (1971), Ethier (1974), Jones and Scheinkinan (1977) and Neary (1985).

The Rybczynski functions $\hat{y}_j(p,l)$ exist as single-valued functions only for the case $m \geqq n$. If all n commodities are produced, and all m factors are fully employed, they satisfy the resource allocation equation

$$B[\hat{w}(p,l)]\hat{y}(p,i) = l. \tag{92}$$

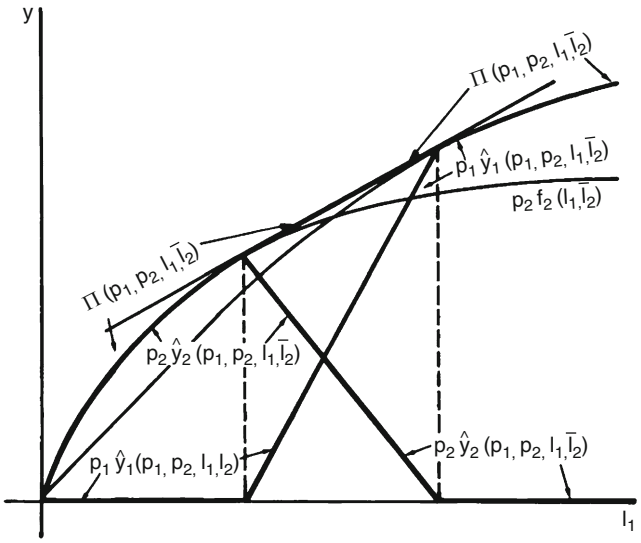
When $m = n$, since then w is locally independent of l, $\hat{y}(p,l)$ is locally linear in l for any fixed p and may be written as

$$\hat{y}(p,l) = B\left[g^{-1}(p)\right] l \tag{93}$$

(cf. Chipman 1971, p. 214, 1972, p. 216). The curious shapes of the Rybczynski functions are illustrated in Fig. 9.

As in the case of the Stolper-Samuelson mapping, one can consider the elasticities of outputs with respect to factor endowments when $m = n$. Denoting $L = \text{diag } l$ and $Y = \text{diag } y$ (not to be confused with the national-income variable Y of

International Trade, Fig. 9



section "The National-Product Function" above), we may define the matrix $\Lambda = L^{-1}BY$ with elements of $\lambda_{ij} = b_{ij}y_j/l_i = \partial \log l_i/\partial \log y_j$ (interpreting l_i in this relationship as requirements or demand for factor i, rather than supply). Its inverse $\Lambda^{-1} = Y^{-1}B^{-1}L$ has elements $\lambda^{ij} = b^{ij}l_j/y_i = \partial \log \hat{y}_i/\log l_j$. From the resource-allocation constraint

$$\sum_{j=1}^{n} b_{ij}y_j = l_i \tag{94}$$

it follows that

$$\sum_{j=1}^{n} \lambda_{ij} = 1,$$

i.e., Λ is row-stochastic (its elements are non-negative and its row sums are equal to unity). By the same reasoning as before, the row sums of Λ^{-1} are equal to unity. In the case $n = m = 2$, adhering to the convention (91) it follows that, when B has positive elements, the off-diagonal elements of Λ^{-1} are negative, and thus its diagonal elements are greater than unity. Thus, $\partial \log \hat{y}_i/\partial \log l_i > 1$ and $\partial \log \hat{y}_i/\partial \log l_i < 0$ for $j \neq i$; in words, a rise in the ith factor endowment will, at given world prices, lead to a more than proportionate rise in the output of the ith commodity, and a fall in the output of the jth commodity ($j \neq i$). As in the case of the Stolper-Samuelson theorem, this obviously does not generalize to higher dimensions unless stronger assumptions are made or weaker conclusions reached. A discussion of the nature of such generalized results will be found in Kemp and Wegge (1969), Wegge and Kemp (1969), Ethier (1974), Jones and Scheinkman (1977) and Neary (1985).

Interindustrial Relationships and other Refinements

The formal model treated so far assumes that production is completely integrated, contrary to fact. Indeed, a large part of international trade is in intermediate products. The main justification for not allowing for intermediate inputs at the very beginning is that it may obscure the logic of the analysis with inessential details. However, in view of the importance of the phenomenon it is desirable at this point to see how the formal framework needs to be modified (cf. McKinnon 1966; Melvin 1969a, 1969b; Khang and Uekawa 1973).

In place of (35) one needs to substitute the production function

$$\begin{aligned} q_j &= f_j(u_{1j}, u_{2j}, \ldots, u_{nj}, v_{1j}, v_{2j}, \ldots, v_{mj}) \\ &= f_j(u_j, v_j) \end{aligned} \tag{95}$$

(assumed homogeneous of degree 1) where q_j denotes gross output of commodity j, and u_{ij} denotes the amount of commodity i used as input to the production of commodity j. Its dual minimum-unit-cost function – equal to the price of commodity j when that commodity is produced – is denoted

$$\begin{aligned} p_j &= g_j(p_1, p_2, \ldots, p_j, w_1, w_2, \ldots, w_m) \\ &= g_j(p, w), \end{aligned} \tag{96}$$

and the input–output and factor–output coefficients are, in accordance with Theorem 5,

$$\begin{aligned} a_{ij}(p, w) &= \partial g_j(p, w)/\partial p_i; \quad b_{ij}(p, w) \\ &= \partial g_j(p, w)/\partial w_i. \end{aligned} \tag{97}$$

The production–possibility set (43) is now replaced by the *net-output-possibility set* defined by

$$y_j = f_i(u_j, v_j) - \sum_{k=1}^{n} u_{jk} \text{ and } \sum_{j=1}^{n} v_j \quad l \tag{98}$$

(cf. Khang and Uekawa 1973). This set is convex. Khang and Uekawa (1973) developed a generalization of Theorem 9 (section "The National-Product Function" above); see also Kemp et al. (1978) and Färe (1979). The concept of a national-product function can also be generalized to this case (cf. Chipman 1985a, pp. 405–6), allowing one to define net supply (Rybczynski) functions.

International Trade and Heterogeneous Firms

Marc J. Melitz

Abstract

Empirical studies of production units within sectors have reported a massive amount of heterogeneity in various performance measures (most notably, size and productivity). This heterogeneity, within sectors, matters for theoretical and empirical models of trade. Trade, or trade liberalization more generally, induces important reallocations between heterogeneous producers in a sector: the smallest, least productive producers are forced to exit, and market shares are further reallocated between less productive producers (who do not export) towards larger, more productive exporters. These reallocations generate a new channel for productivity and welfare gains from trade.

Keywords

Comparative advantage; Export market entry; Extensive and intensive margins of trade; Factor proportions; Firm-level heterogeneity; Firm-level productivity growth; International trade; Intra-industry trade; Market share; Monopolistic competition; Open economy models of growth; Product differentiation; Skill-biased technical change; Sunk costs of entry; Trade costs; Trade liberalization; Trade models

JEL Classifications

F

Census-wide 'micro' level studies of production units for a wide range of countries at all levels of development have documented substantial heterogeneity in virtually all relevant performance measures across these production units. For example, across all US manufacturing plants in 1992, a plant one standard deviation above the mean plant size is 167 per cent bigger, and a plant one standard deviation above the mean plant productivity level (value-added per worker) is 75 per cent more productive (Bernard et al. 2003). (More precisely, the standard deviation of log sales is 1.67 and that of labour productivity is 0.75.) These represent massive differences in performance outcomes, which are also reflected in differences in other key plant characteristics. Furthermore, the extent of this heterogeneity does not diminish much when looking within narrowly defined sectors. In the case of the US plants, the 75 per cent productivity difference mentioned above only drops to a 66 per cent difference when controlling for productivity differences across more than 400 different sectors.

These large differences in firm performance are also strongly correlated with the firm decision to engage in international transactions (such as exporting, importing intermediate goods from foreign suppliers, or investing in foreign subsidiaries): only a small proportion of firms report any such activities, even within narrowly defined sectors; and those firms are substantially larger and more productive than their counterparts with no international contacts in the same sector. (This pattern has been documented at both the firm and the plant level for a very wide range of countries. From here on out, I will mostly focus on differences between exporting and non-exporting firms, although similar differences have also been documented concerning multinational firms and firms that import intermediate goods from foreign suppliers.) For the United States, Bernard et al. (2006a, b, c) report that manufacturing plants are more than twice as large (value of shipments) as and 14 per cent more productive (value-added per worker) than their non-exporting counterparts in the same sector. (Bernard et al. 2006a, b, c, provide an extensive description of firm-level differences related to international trade based on US manufacturing data and also survey the related empirical and theoretical literatures.) Bernard et al. (2006a, b, c) also report how these exporting firms exhibit other different characteristics relative to non-exporters: they are more capital- and skill-intensive, and pay higher wages.

This strong correlation between export status and firm characteristics (notably higher productivity) naturally leads to the follow-up question of causality. A very large number of studies have examined this question, usually focusing on a firm's productivity trajectory over time relative to its export market entry decision. Virtually all these studies find a strong self-selection effect: firms are relatively more productive prior to their entry into export markets. (Two early influential papers in this area are Bernard and Jensen 1999, and Clerides et al. 1998). Several of these studies further reject the hypothesis of firm-level productivity growth following export market entry, although some studies, especially for developing countries, do report such a link. (See, for instance, Loecker 2007; Topalova 2004; Biesebroeck 2005; and the survey by Girma et al. 2004). However, this distinction – based on the timing of the export market entry – has been blurred given the evidence from some recent studies that firms make innovation/technology use decisions based on current or anticipated export market participation, as highlighted by Bustos (2006), Verhoogen (2007), and Trefler and Lileeva (2007). In such a case, productivity and exporting decisions are both endogenous with respect to one another, and the timing of the export market entry can no longer be used to identify causality. (Yeaple 2005, theoretically studies this joint technology adoption and export decision by firms, and explores the consequences for the return to skill – highlighting how skill-biased technological change may be induced by trade.)

Nonetheless, the results obtained clearly indicate that it is initially more successful firms that make the joint decisions concerning innovation (or 'higher' technology use) and export status. In other words, the least successful firms overwhelmingly tend to undertake neither activity.

Another part of the recent empirical literature using micro-level data has examined the consequences of this link between export status and productivity when the exposure to trade is changing (predominantly because trade costs are decreasing over time). In such a case, trade liberalization induces some reallocations between exporters and non-exporters competing in the same sectors (see Tybout 2003, for a survey of this literature). One influential such study by Pavcnik (2002) finds that most of the 25 per cent productivity increase in export-competing sectors in Chile between 1979 and 1986 is explained by reallocations between producers (generated by entry, exit, export market entry, and market share reallocations). However, since significant changes in trade regimes are also part of a larger set of substantial macroeconomic changes for the involved countries (as was the case for Chile), it nevertheless remains difficult to associate this type of reallocation-induced productivity growth to the direct effects of trade liberalization. One notable exception is Bernard et al. (2006c), who show that reductions in trade costs for US plants substantially increase both the probability of exit and that of exporting among non-exporters. Given the productivity advantage of exporters, this induces reallocations in favour of the more productive exporting plants and hence increases average industry productivity (which is also confirmed by Bernard et al. 2006a, b, c, as a result of the decrease in trade costs).

Clearly, these empirical patterns cannot be addressed by trade models based on representative firms. Such models, by construction, predict that trade affects all firms in a sector in similar ways. (Note that extensive firm-level heterogeneity per se is not necessarily problematic for a representative firm model of trade so long as firms, on average, respond in similar ways to trade. However, the evidence reviewed clearly shows that this is not the case.) In response to this empirical evidence, theoretical models of trade have been developed to incorporate firm-level productivity differences, and analyse the consequences for the effects of trade liberalization. One class of models, developed by Bernard et al. (2003) and Eaton and Kortum (2008), introduce stochastic firm productivity into the multi-country Ricardian model analysed in Eaton and Kortum (2002). In this class of models, there is a fixed number of products that can be produced by competing firms in all countries. All these firms (both in the same country but also across countries) use different technologies to produce the same good (based on a stochastic productivity

draw) – hence the Ricardian framework. Consumers in any given country buy each good from the lowest-cost producer across all countries. Due to trade costs, several firms producing the same good can survive if they are located in different countries (although each firm is the sole supplier to any given destination). This model thus emphasizes the resulting competition between firms to be this exclusive supplier. Bernard et al. (2003) show how such a model can be calibrated to fit both micro-level data on US producers and macro-level data on cross-country trade and aggregate production across countries. The calibrated model can then be used to analyse many counterfactual predictions involving the consequences of trade liberalization.

Another class of models developed in Melitz (2003) and Melitz and Ottaviano (2005) eschews the analysis of the direct competition between firms to produce the same good by using a monopolistic competition framework: each firm produces its own distinctive differentiated good. These models incorporate firm heterogeneity into the one-sector models of intra-industry trade (the 'new' trade theory) developed in Krugman (1979, 1980). In this type of model, the product variety available to consumers in any given country varies endogenously with the characteristics of the country and the trade costs linking it to its trading partners (these affect the endogenous number of varieties produced domestically, as well as the endogenous fraction of firms from all trading partners that export to that country). Firms face sunk costs of entry, along with uncertainty concerning their future productivity (or also possibly the quality of the differentiated good that is under development). Upon entry, each firm instantaneously learns about its productivity level, modelled as a draw from a known distribution. Due to the sunk nature of the entry costs, firms with heterogeneous productivity levels remain active and produce. The least productive firms face negative profits and therefore exit. As exporting is costly, only the relatively more productive firms (among those surviving) choose to export, while the remaining firms only serve their domestic market. Exporting is not profitable for these firms, either because it involves fixed or sunk costs, or because import demand is driven to zero at prices below the firms' delivered cost.

Both classes of models predict that trade liberalization induces the type of reallocations between firms that was previously described: the least productive firms are constrained to exit, new firms enter the export market, and market shares are reallocated towards more productive firms. These reallocations generate both aggregate productivity and welfare gains. Both classes of models also predict an important empirical regularity regarding bilateral trade flows: that differences in these trade flows reflect both differences in the amount of each good traded (the intensive margin of trade) as well as differences in the number of goods traded (the extensive margin of trade). (See Bernard et al. 2005; Broda and Weinstein 2006; Broda et al. 2006; Eaton et al. 2004; and Kehoe and Ruhl 2003, for some empirical applications). Helpman et al. (2007) and Chaney (2006) show how the framework of Melitz (2003) can be extended to derive a gravity specification for bilateral trade flows where trade costs affect both the extensive and intensive margins of trade. Both papers highlight the empirical importance of incorporating changes in trade at both margins.

Due to the absence of strategic interactions between firms, the monopolistic competition model of Melitz (2003) provides a convenient framework for the modelling of additional firm-level decisions in an open economy environment – where heterogeneous firms self-select into different types of activities. This framework can thus also explain why only a fraction of firms choose to become multinationals and operate foreign affiliates (horizontal foreign direct investment, FDI) as in Helpman et al. (2004) or integrate with their foreign suppliers (vertical FDI) as in Antras and Helpman (2004). (Helpman 2006, provides a much more extensive review of the related models.) Additionally, other firm-level decisions that are also affected by the exposure to international trade can be incorporated: the choice of technology as in Acemoglu et al. (2007), the level of investment in innovation as in Atkeson and Burstein (2006), or the range of products produced and exported

within multi-product firms as in Bernard et al. (2006b). Lastly, the structure from Melitz (2003) has also been fruitfully integrated into various other types of models that rely on the basic monopolistic competition of trade. This includes extension to two-sector models of trade with comparative advantage and factor proportion differences (Bernard et al. 2007), open economy models of growth (Baldwin and Robert-Nicoud 2006), and international macro-dynamics (Ghironi and Melitz 2005). In each case, the addition of firm-level heterogeneity allows the models to explore additional important features upon which a model with representative firms remains silent.

See Also

- ▶ International Trade Theory
- ▶ International Trade, Empirical Approaches to
- ▶ New Economic Geography
- ▶ Ricardian Trade Theory
- ▶ Trade, Technology Diffusion and Growth

Bibliography

Acemoglu, D., P. Antras, and E. Helpman. 2007. Contracts and technology adoption. *American Economic Review* 97: 916–943.

Antras, P., and E. Helpman. 2004. Global sourcing. *Journal of Political Economy* 112: 552–580.

Atkeson, A., and A. Burstein. 2006. *Trade costs, pricing to market, and international relative prices*. UCLA: Mimeo.

Baldwin, R.E., and F. Robert-Nicoud. 2006. Trade and growth with heterogenous firms. Working paper no 12326. Cambridge, MA: NBER.

Bernard, A.B.., and J.B. Jensen. 1999. Exceptional exporter performance: Cause, effect, or both? *Journal of International Economics* 47: 1–25.

Bernard, A.B.., J. Eaton, J.B. Jensen, and S. Kortum. 2003. Plants and productivity in international trade. *American Economic Review* 93: 1268–1290.

Bernard, A.B.., J.B. Jensen, and P.K. Schott. 2005. Importers, exporters, and multinationals: A portrait of firms in the US that trade goods. Working paper no. 11404. Cambridge, MA: NBER.

Bernard, A.B.., J.B. Jensen, S. Redding, and P.K. Schott. 2006a. Firms in international trade. Working paper no. 13054. Cambridge, MA: NBER.

Bernard, A.B.., S. Redding, and P.K. Schott. 2006b. Multi product firms and trade liberalization. Working Paper No. 12782. Cambridge, MA: NBER.

Bernard, A.B.., J.B. Jensen, and P.K. Schott. 2006c. Trade costs, firms and productivity. *Journal of Monetary Economics* 53: 917–937.

Bernard, A.B.., S. Redding, and P.K. Schott. 2007. Comparative advantage and heterogeneous firms. *Review of Economic Studies* 74: 31–66.

Biesebroeck Van, J. 2005. Exporting raises productivity in sub-Saharan African manufacturing firms. *Journal of International Economics* 67: 373–391.

Broda, C., and D.E. Weinstein. 2006. Globalization and the gains from variety. *Quarterly Journal of Economics* 121: 541–585.

Broda, C., J. Greenfield, and D. Weinstein. 2006. From groundnuts to globalization: A structural estimate of trade and growth. Working paper no. 12512. Cambridge, MA: NBER.

Bustos, P. 2006. *Rising wage inequality in the Argentinean manufacturing sector: The impact of trade and foreign investment on technology and skill upgrading*. Mimeo: CREI, Universitat Pompeu Fabra.

Chaney, T. 2006. *Distorted gravity: Heterogeneous firms, market structure and the geography of international trade*. Mimeo: University of Chicago.

Clerides, S.K., S. Lach, and J.R. Tybout. 1998. Is learning by exporting important? Micro dynamic evidence from Colombia, Mexico, and Morocco. *Quarterly Journal of Economics* 113: 903–947.

Eaton, J., and S. Kortum. 2002. Technology, geography, and trade. *Econometrica* 70: 1741–1779.

Eaton, J., and S. Kortum. 2008. *Technology in the global economy: A framework for quantitative analysis*. Princeton: Princeton University Press.

Eaton, J., S. Kortum, and F. Kramarz. 2004. Dissecting trade: Firms, industries, and export destinations. *American Economic Review* 94: 150–154.

Ghironi, F., and M.J. Melitz. 2005. International trade and macroeconomic dynamics with heterogeneous firms. *Quarterly Journal of Economics* 120: 865–915.

Girma, S., D. Greenaway, and R. Kneller. 2004. Does exporting increase productivity? A micro-econometric analysis of matched firms. *Review of International Economics* 12: 855–866.

Helpman, E. 2006. Trade, FDI, and the organization of firms. *Journal of Economic Literature* 44: 589–630.

Helpman, E., M.J. Melitz, and S.R. Yeaple. 2004. Export versus FDI with heterogeneous firms. *American Economic Review* 94: 300–316.

Helpman, E., M.J. Melitz., and Y. Rubinstein. 2007. Estimating trade flows: Trading partners and trading volumes. Working paper no. 12927. Cambridge, MA: NBER.

Kehoe, T.J., and K.J. Ruhl. 2003. *How important is the new goods margin in international trade?* Mimeo.

Krugman, P.R. 1979. Increasing returns, monopolistic competition, and international trade. *Journal of International Economics* 9: 469–479.

Krugman, P.R. 1980. Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70: 950–959.
Loecker De, J. 2007. Do exports generate higher productivity? Evidence from Slovenia. *Journal of International Economics* (forthcoming).
Melitz, M.J. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71: 1695–1725.
Melitz, M.J. and G.I.P. Ottaviano. 2005. Market size, trade, and productivity. Working paper no. 11393. Cambridge, MA: NBER.
Pavcnik, N. 2002. Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants. *Review of Economic Studies* 69: 245–276.
Topalova, P. 2004. Trade liberalization and firm productivity: The case of India. Working paper no. 04/28, International Monetary Fund.
Trefler, D., and A. Lileeva. 2007. *Does improved market access raise plant-level productivity?* Mimeo.
Tybout, J.R. 2003. Plant- and firm-level evidence on 'new trade theories'. In *Handbook of international trade*, ed. E.K. Choi and J. Harrigan. Oxford: Basil Blackwell.
Verhoogen, E. 2007. *Trade, quality upgrading and wage inequality in the Mexican manufacturing sector: Theory and evidence from an exchange-rate shock.* Mimeo.
Yeaple, S.R. 2005. Firm heterogeneity, international trade, and wages. *Journal of International Economics* 65: 1–20.

International Trade Theory

James E. Anderson

Abstract

International trade theory provides explanations for the pattern of international trade and the distribution of the gains from trade. The theory convinces most economists of the benefits of liberal trade. But many non-economists oppose liberal trade. Opponents include some who may have encountered trade theory but nevertheless fall prey to fallacious reasoning. This article attempts to convey why trade theory is so persuasive to economists and also to deal with why many non-economists are not persuaded.

Keywords

Absolute advantage; Adjustment costs; Autarky; Bilateral trade; Comparative advantage; Comparative labour productivity advantage; Competitive advantage; Consumption taxation; Economic theory of gravity; Economies of scale; Endogenous advantage; External budget constraint; Factor immobility; Factor proportions theory; Free trade; Gains from trade; General equilibrium trade model; Heckscher–Ohlin trade theory; Imperfect competition; Information limitations; International trade theory; Lump-sum transfers; Monopolistic competition; Multilateral resistance; Multiple equilibria; Partial equilibrium; Product variety; Productivity shocks; Protection; Ricardian trade theory; Substitution effect

JEL Classifications

F1

Why do nations trade the products they do? Is trade a good thing? The theory of international trade provides answers. The answers are both convincing and elegant; hence, the vast majority of economists agree about the desirability of liberal trade. But the argument is also subtle and often misunderstood or distorted. Thus, a large proportion of the general population tends to oppose liberal trade from confusion. This article attempts to convey why the answers convince most economists and why their liberal trade position is so often misunderstood. The article's focus is theory, but theory convinces when it succeeds in fitting the data. Thus, passing reference will be made to empirical findings, a sensibility much more thoroughly developed in the graduate textbook of Feenstra (2003).

'Buy low, sell high' logic leads economists to comparative advantage theory. Comparative advantage means the comparison of *relative* price differences between nations to explain the pattern of trade. For example, compare the relative price of wheat in terms of cheese at home with the same relative price in the foreign economy in a hypothetical equilibrium with no trade (autarky) or with restricted trade. The country with the lower relative price of wheat is said to have a comparative advantage in wheat while the other country has, symmetrically, a comparative

advantage in cheese. Buy low, sell high logic predicts that a country will export the good in which it has a comparative advantage. (In the case of many goods, the prediction is that a country will on average export goods which are relatively cheap in the absence of trade and import goods which are relatively expensive in the absence of trade. The prediction is about correlation. Bernhofen and Brown 2004, show that Japan's opening to trade in the 1850s reveals data consistent with the prediction.)

Notice that the focus on relative prices tends to cancel out forces (exchange rate manipulations, environmental or labour standards) which cause national differences in levels of non-traded factor (or goods) prices. Note also that by this reasoning a country *must* have a comparative advantage in some good. Prices of non-traded factors of production adjust in general equilibrium so that each country ends up in the trade equilibrium with a competitive or absolute cost advantage in the good in which it has a comparative advantage. Partial equilibrium thinking takes factor prices as given and does not impose the external budget constraint that requires exports to pay for imports. Partial equilibrium reasoning leads to misunderstandings explored below as the absolute advantage fallacy.

Comparative advantage differences between nations are explained by exogenous differences in national characteristics. Labour differs in its productivity internationally and different goods have different labour requirements, so comparative labour productivity advantage was Ricardo's predictor of trade patterns. Ricardian trade theory is useful in its simplicity and even rather loosely confirmed by empirical evidence. The factor proportions theory added relative factor endowment differences to the exogenous explanation of comparative advantage (Jones 1987). More capital-abundant countries have higher labour productivity, but the advantage gained relative to the less abundant countries varies with the relative capital intensity of the good's technology. Combining technology and endowment differences appears to account well for actual trade patterns (Davis and Weinstein 2002).

Trade theory also encompasses endogenous differences between countries. One focus is on economies of scale. The wider market due to trade induces a cost advantage in an industry in one of the countries. Another theory is based on monopolistic competition, whereby the wider markets due to trade increase product variety as buyers seek the special characteristics of foreign brands. Differentiated products trade flows both ways within product categories.

Trade costs also shape the pattern of trade. The economic theory of gravity explains the complex bilateral trade patterns among countries. Actual trade is much lower than gravity predicts in a frictionless world, providing evidence of trade costs much larger than those due to policy or transportation. The costs are well explained by geography and a set of national differences. The stability of the relationships over time suggests that these costs change slowly.

There are gains from trade in all these models. But the division of the gains will be uneven and there will be losers. Distribution matters in two ways, between and within nations. Internationally, with only mild qualifications, gains are shared between nations: some trade is better than none. Each nation can act through trade policy to take more of the gain, however, leading to destructive trade wars with mutual losses. Within national economies, there are gains on average but there are ordinarily losers. National institutions act to redistribute some of the gains (US Trade Adjustment Assistance) or provide temporary relief from losses due to trade (escape clause protection), at the cost of lowering the overall gain from trade.

The topics of this outline are developed below in more detail. Section 1 examines the causes of comparative advantage. Section 2 exposes the absolute advantage fallacy. Section 3 reviews endogenous advantage. Section 4 sets out the economic theory of gravity and its implications. The concluding section examines the gains from trade.

Comparative Advantage

Ricardo explained comparative advantage as due to differences in labour productivity. Suppose that it takes two hours of labour to produce a bushel of

wheat in the home country, while it takes four hours of labour to produce a bushel of wheat in the foreign country. Also, it takes three hours of labour to produce a pound of cheese in the home country while it takes eight hours of labour to produce a pound of cheese in the foreign country.

Ricardo saw that the world trade equilibrium would result in the home country exporting cheese and the foreign country exporting wheat. This is because, in the absence of trade, a pound of cheese is worth 1.5 bushels of wheat (three hours per pound of cheese divided by two hours per bushel of wheat) in the home country while a pound of cheese is worth two bushels of wheat in the foreign country. The labour market equilibrium which accompanies such a trade equilibrium must have a foreign wage of at most one-half of the home wage (since, with a foreign wage equal to one-half the home wage, a bushel of wheat costs the same amount in each country, allowing production in both). If we consider a low-wage foreign economy, the labour market equilibrium accompanying the trade equilibrium could have a foreign wage no lower than three-eighths of the home wage (since in this case a pound of cheese costs the same amount in each country).

Notice that countries export the good in which they have the comparative labour productivity advantage, cheese for the home country and wheat for the foreign country. The numbers chosen make no difference to the logic; what is essential is that comparative labour productivities differ. One special aspect of the numbers deserves emphasis, however: the home country has an absolute labour productivity advantage in both goods yet trade occurs regardless.

Subsequent developments of trade theory generalized the production model. The essence of comparative advantage theory remains: trade is due to differences in relative prices that would obtain in the absence of trade, and an average of each country's citizens gain from such trade. The Heckscher–Ohlin analysis of the factor proportions model predicted that a country would have a comparative advantage in the good which made relatively intensive use of its relatively abundant factor. Thus, if the home country were relatively abundant in capital (which would explain why its labour was so much more productive in the preceding example), it would have a comparative advantage in the good which used capital relatively intensively (cheese in the preceding example). Conversely, the foreign country is relatively abundant in labour and has a comparative advantage in the good which uses labour relatively intensively (wheat in the example above).

Trade in goods compensates for the international immobility of factors. The factor content extension of Heckscher–Ohlin trade theory predicts that trade patterns permit each country to consume factor services as if it were in a completely integrated world, smoothing out differences in national factor endowments. Recent empirical work has met with striking success in combining factor endowment differences with technology differences as an explanation of observed trade patterns (Davis and Weinstein 2002).

Comparative advantage theory is much more general than the preceding discussion of special cases (Deardorff 1984), but predictions about the pattern of trade weaken with generality. On average a country will import goods that would be relatively expensive in the absence of trade. See the Appendix for a technical statement. See Bernhofen and Brown (2004) for confirming evidence based on Japan's opening to trade in the 1850s. The assumptions of the general model are that (*a*) price-taking consumers minimize the expenditure needed to realize any level of utility (real income), and (*b*) producers behave so as to maximize the national product given the resource endowments. Assumption (*a*) implies downward-sloping demand curves in the generalized form. Assumption (b) leads to upward-sloping supply curves in the generalized form. Scale economies and imperfect competition, treated below in the section on endogenous advantage, can lead to the violation of assumption (*b*).

The Absolute Advantage Fallacy

Businessmen naturally compare the money cost of the same good in different locations to draw inferences about the direction of trade. Absolute cost

advantage appears to imply that a nation imports goods that are cheaper abroad and exports goods that are more expensive abroad. The reasoning is insidious because it makes sense in many contexts. Absolute advantage appropriately addresses the householder's question of which good should be purchased, the businessman's question of how tough his competitors are. The individual businessman can appropriately take all other prices as given when contemplating his own actions, such as entering a new export market.

To see the difference between absolute and comparative advantage reasoning clearly, return to the Ricardian example above. If wages (measured in a common currency) were equal in the two countries prior to the opening of trade, the home country would have a 'competitive' or absolute advantage in both goods: it could undersell the foreign country in both wheat and cheese. Foreign businessmen would naturally be worried that they would all be driven from the market. This universal bankruptcy could not be an equilibrium, however, because the foreign workers would have no income to pay for home-produced goods. The imbalance between expenditure and income would also mirror the absence of exports to pay for imports. Market equilibrium would be reached through price changes, lowering the foreign wage or raising the home wage until the foreign workers could be employed in the industry in which the foreign economy has the comparative advantage. (Unless the two currencies were pegged, the exchange rate of the foreign economy could depreciate and create the same effect.) More general models of production lead to the same conclusion: *equilibrium costs will adjust to confer absolute advantage in the good in which each country has a comparative advantage.*

With many goods, comparative advantage applies to ranges of goods rather than to a single good, and the dividing line between comparative advantage and disadvantage is endogenous. The absolute advantage is weak in the mathematical sense in the case where both countries continue to produce the good.

Another illustration of the absolute advantage fallacy arises in popular concerns about the rapid productivity growth of China compared with that of the United States. A ten per cent improvement in productivity will indeed secure a ten per cent cost advantage for the businessman over his competitor. A ten per cent improvement in all Chinese productivity relative to the United States is unlikely to change comparative advantage (indeed, in the Ricardian example, comparative labour productivity advantage is unchanged) because Chinese wages will rise relative to US wages. Similarly, a ten per cent drop in all US productivity due to tighter environmental regulations will be unlikely to change comparative advantage because US factor returns will fall.

The widespread practice of making international comparisons of 'competitive advantage' is essentially misguided because it suggests the metaphor of a race. The race metaphor is extended in concerns about 'a race to the bottom', which supposedly expresses the dilemma of countries seeking to implement pollution or labour standards but being pressured to lower standards by their competition with foreign countries that have low standards. But nations do not 'compete' as firms do. A firm may well be unable to survive after implementing pollution reduction when its competitors abroad do not follow suit and no other prices change in the new equilibrium. Nations cannot similarly put themselves out of business because factor prices will change in the new equilibrium. Polluting industries may or may not survive at the new factor prices under the new regulations, but the nation's factors will be productively employed somewhere in the economy. Pollution reduction is costly with or without trade; nothing about the nature of a trading economy makes any essential difference to the nation's ability to implement desired standards. The desirability of trade is an *essentially* separate matter.

Endogenous Advantage

Many goods are traded because they are simply unavailable from local production. Some kinds of availability are exogenous to the interaction of nations – diamonds and oil are found only in a few locations. Endogenous availability is in

contrast driven by advantage arising from the economic interaction of nations. Endogenous advantage normally coexists with comparative advantage but it is simpler to consider special cases independent of comparative advantage. Theory focuses on endogenous advantage resulting from economies of scale. (In a formal but trivial sense, oil or diamond trade can be seen as comparative advantage trade – big oil deposits lead to a low relative price of oil where they are found. Moreover, comparative advantage trade is often associated with the disappearance of some industries in some countries. Neither of these associations of comparative advantage with availability is essential to the model, however.)

Trade based on scale economies features the possibility of multiple equilibria – one country will produce a good with scale economies but which nation ends up producing it can be a matter of chance. Since advantage is endogenous, it appears attractive in developing countries to attempt to reverse the historical head start of rich countries by starting up production behind protection and then later being able to compete on world markets. The record of success in such efforts is mixed.

Openness to trade will generally allow economies of scale to be more thoroughly exploited, so this is a new source of gains from trade. Moreover, wider markets may support a wider range of products, still another source of gains from trade. Each country shares in the gains from trade with scale economies under conditions that appear to be met in practice. (This claim is based on the results from numerous simulation models of trading economies that have been developed since the mid-1970s.) The theoretical possibility that a country can lose from trade based on scale economies has drawn a lot of attention from development economists in particular (Ethier 1982b). (Losses result when a trading equilibrium has a country importing the good with scale economies while still producing it. Since domestic scale is smaller, unit costs are higher, meaning that market forces perversely 'choose' to import a good with higher price than in autarky. Simulation models have not found such equilibria but they are possible.) Gains can be guaranteed if a country expands production in goods with scale economies, so it looks more attractive to use policy to promote production of such goods.

Scale economies come in two forms: external to the firm and internal to the firm. External scale economies are typified by specialized labour markets such as Silicon Valley, where the concentration of the market reduces search costs for computer engineers. External scale economies need not be location-specific, however. Increases in the scale of downstream final production can permit carrying on upstream input production with a specialized process that is cheaper at large enough scale. Such scale economies can operate at the level of the world economy and appear to be bound up with the recent phenomenon of outsourcing (Ethier 1982a). Global scale economies tend to guarantee mutual gains from trade among countries.

Internal scale economies are associated with imperfect competition when the size of the firm looms large relative to the market size. Trade tends to intensify competition and thus to reduce the inefficiency of monopoly, another gain from trade.

The most fruitful form of imperfect competition for trade theory has been monopolistic competition. Only Ford Motor Co. produces Ford autos (monopoly) but dozens of brands compete for auto buyers. Each design has a fixed cost of design (and marketing) which must be covered by sales net of variable cost. The total market size limits the number of designs which can profitably be produced. A signal accomplishment of trade theory in the 1980s was the embedding of monopolistic competition in a general equilibrium trade model (Helpman and Krugman 1985; Ethier 1982a). Progress was enabled by the simplifying assumption of symmetric firms: all brands were equally desirable and all firms' costs were the same.

Monopolistic competition provides an explanation of the two-way international trade that is found in many products such as autos, and of why two-way trade is more prevalent between similar countries. Trade between rich and poor countries, in contrast, is explained mainly by comparative advantage as autos exchange for agriculture.

Relative country size matters too, the home-market effect of Krugman (1980). Here the insight has been rigorously proved only for a two-country example. Start with two equally sized countries, then increase one relative to the other. Trade costs imply that the larger country will have a more than proportionally larger share of brands. Intuitively, with access to foreign markets being costly the home market, being larger, allows scale economies to be more readily exploited, increasing the larger country's share of differentiated goods production more than its share of world income.

Monopolistic competition theory has recently focused on the heterogeneity of firms. If the symmetry of firms on the demand side is retained, differences in firms' productivities imply differential responses to trade. The best firms export disproportionately while imports drive out the worst firms. Fixed trade costs add explanatory power; only the best firms choose to incur the cost of trade. A key element of the model is productivity shocks, firms discover their productivity after committing fixed costs. The distribution of surviving firms is related to the distribution of productivity shocks as well as economic determinants. The models of Bernard et al. (2003) and Melitz (2003) deserve special attention. The former focuses on competition within a variety while the latter focuses on competition across varieties. Both models imply new gains from trade in the form of overall productivity gains: opening trade causes the exit of weak firms and the expansion of strong ones.

Bilateral Trade Patterns

The trade theories presented above are focused on explaining the cross-commodity trade pattern of essentially two trading countries. The contemporary world of more than 100 countries (most of which are collections of distinct economic regions) has complex trade patterns.

The economic theory of gravity complements the preceding models by providing an explanation of bilateral trade (Anderson and van Wincoop 2004). Gravity fits the data well and reveals important information. The model is based on four assumptions: expenditure on goods from all sources is equal to income from sales to all sources, markets for all goods clear, (more restrictively) each country or region produces a unique good, and all countries have the same tastes for goods.

The third assumption – products differentiated by place of origin – appears to be the most restrictive. In practice, only models of this type do at all well in fitting bilateral trade patterns. Monopolistic competition provides one explanation for why products appear to be differentiated by place of origin. Eaton and Kortum (2002) show alternatively that productivity shocks in a Ricardian model will select producers within product lines, resulting at the aggregate level in what appears to be two-way trade. In either case, gravity ends up describing trade flows.

In a frictionless world, gravity theory predicts that the bilateral trade in a commodity as a share of world production of the commodity will be equal to the product of the source country's share of world production of the commodity times the consuming country's share of expenditure on the commodity. Alternatively, the model predicts that size-adjusted trade, the bilateral flow divided by the product of source country supply and consuming country expenditure, should be constant across country pairs in a frictionless world.

Actual trade flows are far smaller than the frictionless prediction (while shipments within regions are far larger, home bias). The deviations of actual bilateral trade from the frictionless prediction allows inference about bilateral trade costs. Distance appears to be more costly than can be accounted for by transport costs. Other costs are associated with non-contiguity, language barriers, exchange rate barriers, insecurity and other plausible bilateral characteristics. Just crossing a border imposes a cost which is larger than can be explained by policy variables.

Trade flows in the model are predicted to vary with relative resistance, equal to the ratio of the direct bilateral trade cost to the product of inward and outward multilateral resistance. Multilateral resistance is an index of bilateral trade costs, inward from every source to a particular destination or outward from a particular source to every destination. Multilateral resistance is linked to

country size and thus to explaining an important aspect of trade patterns. Since borders are costly, a big country tends to have lower multilateral resistance than does a small country because a smaller fraction of its shipments must cross borders. The size-adjusted internal trade of big countries will be smaller than that of small countries because big countries have higher relative resistance to their internal trade. These differences can be quite dramatic, as shown by studies of US and Canadian trade (Anderson and van Wincoop 2003), where the United States is about ten times larger than Canada.

Division of the Gains

Professional economists generally support liberal trade because theory and evidence persuade them that there are gains from trade in an average sense in all these models of the determinants of trade. But the division of the gains will be uneven and there can be losers. Most policy intervention with trade is explained by the policymakers' desire to alter the distribution of gains.

The gains from trade reasoning is illustrated with comparative advantage-based trade. Focus on a 'typical' household. Suppose that in autarky equilibrium, as in the Ricardian numerical example, the Home typical householder is willing to swap 1 unit of cheese for 1.5 units of wheat. That is, he would be indifferent to moving his consumption and production a small distance to offer the market 1 cheese for 1.5 wheat or 1.5 wheat for 1 cheese. Suppose that a typical foreign country household in the autarky equilibrium is willing to swap 2 wheat for 1 cheese. Now allow frictionless trade, and suppose for illustrative purposes that the new equilibrium price is equal to 1.75 wheat per unit of cheese. (Generally the price must lie between 1.5 and 2, always implying mutual gains.) Each Home household offers cheese to Foreign households in exchange for their wheat. Formerly it cost 2 wheat for 1 cheese in Foreign but now the 2 wheat will procure 1 cheese and leave 0.25 wheat left over, a gain from trade. Similarly, each Home household can obtain 1.75 wheat for 1 cheese where formerly this would procure only 1.5 wheat, a gain from trade of 0.25 wheat. Both households and hence both nations gain from trade. The numbers chosen illustrate a general principle: mutual gains result from trade when autarky relative prices differ. See the Appendix for a more formal discussion.

The mutual gains from trade claim may seem dubious because, with the numbers chosen, trade equilibrium requires that foreign wages must be lower than home wages. In effect, trade facilitates an exchange in which more than one unit of foreign labour exchanges for one unit of home labour – the home country is 'exploiting' foreign labour. Some anti-trade sentiment on the left in rich countries is based on this observation. (Marxism embeds the observation in a wider system of analysis, but it probably is no longer a basis for much sentiment on the left.) Nevertheless, foreign labour gains from trade, as does home labour. Prior to trade, a pound of cheese cost 1.5 bushels of wheat in the home country while it cost 2 bushels of wheat in the foreign country. By specializing in wheat production and exchanging it for cheese, foreign workers can obtain cheese more cheaply, at a price somewhere between 2 and 1.5 bushels of wheat. This exchange must make them better off. As for home workers, prior to trade, a pound of cheese obtained 1.5 bushels while with trade it obtains somewhere between 1.5 and 2 bushels. This must make home workers better off. Concern about the 'fairness' of the exchange in rich countries should lead to policies which might actually help the poor countries. Trade theory shows that anti-trade policies by rich countries will instead on average harm the poor countries.

Scale economies and imperfect competition models of trade suggest further gains. With scale economies, trade implies that the force of wider markets drives costs lower. With imperfect competition, trade stimulates competition and drives profit margins lower. Trade equilibrium with monopolistic competition suggests that consumers and intermediate input users gain from more variety of differentiated products (see Helpman and Krugman 1985).

The distribution of the gains matters, both between and within nations. Nationalist trade

policy can take more of the gains, leading to destructive trade wars with mutual losses. Negotiation of trade agreements and their enforcement through international institutions such as the World Trade Organization (WTO) help to restrain the destructive tendencies of unilateral action. Nations have an incentive to participate in negotiations and to join institutions such as the WTO because some trade is better than none for each nation. Theoretical qualifications to this statement must be entered in models of trade involving scale economies and imperfect competition, but in practice simulation of such models suggests that some trade remains better than none for each nation.

Within national economies the division of gains issue is much sharper: some members of a nation ordinarily lose from trade. Ricardo's one-factor trade model submerges income distribution. Multi-factor production models feature groups who must lose from trade. Loosely speaking, these groups are associated with import competing production (see Jones 1987).

In equilibrium the gains must ordinarily outweigh the losses within each nation, by the preceding national-gains-from-trade argument. For an economy with nonidentical households, this implies that there are gains from trade on average. Under special circumstances the gains can be redistributed so that all households gain. In practice, these circumstances are rarely met completely. Even so, most economists tend to favour efficiency-enhancing policies such as liberal trade on the pragmatic grounds that efficiency-reducing policies such as protection also cause gainers and losers, so it is better to go with the larger net gains and supplement them with feasible programmes to compensate the most obvious losers.

(A benevolent and very powerful government can in principle calculate and implement the lump-sum transfers – negative for gainers, positive for losers – that are required to achieve redistribution so that all gain. In practice, information is more limited and implementation more difficult – because households modify their behaviour to reduce their tax or increase their subsidy – than with the lump-sum story. Trade and public economic theory have relaxed the conditions somewhat. Income taxation can in some circumstances achieve redistribution with efficiency, but information limitations rule them out as a practical matter; see Guesnerie 2001. Dixit and Norman 1986, show that a system of consumption taxes – differentially taxing each commodity – that sacrifices some of the gains from trade is powerful in achieving gains for all. For a qualification of their argument, see Kemp and Wan 1986. Again, information limitations vitiate the applicability of this idea. Finally, a government that can discriminate powerfully between households is sure to be lobbied intensively by those able to organize politically, to the detriment of the unorganized.)

What if losers are not compensated? A person taking this question seriously must decide on liberal trade by weighting individual gains and losses. Ethical considerations give more weight to losses or gains to the poor than to the rich. The case for liberal trade is strengthened by ethical considerations because the illiberal trade policies of rich countries hurt the poor disproportionately, as documented by Gresser (2002). Poor countries have comparative advantage based on cheap low- skilled labour, hence discrimination against their exports harms the poor citizens of poor countries. At home in rich countries, protection makes food and clothing more expensive, a regressive tax on poor consumers. Among the poor, losers from protection appear almost surely to outweigh gainers.

On the way to equilibrium, it is theoretically possible that adjustment cost losses may temporarily exceed gains, justifying temporary relief measures. For example, workers displaced by import competition may be unemployed for a time. Extensive investigation of US cases suggests that such adjustment cost losses from trade are small, of short duration, and are swamped by the gains from trade. A typical investigation reports that the net cost to the economy of using protection to reemploy a worker far exceeds the wage the worker would receive in the job, usually several (up to ten) times the wage. In practice, therefore, temporary protection for workers cannot be justified on efficiency grounds, though it remains possible to justify it on equity grounds.

Economists in favour of liberal trade point out that protection can be replaced with much less inefficient methods of compensation to displaced workers.

A substantial part of the opposition to liberal trade is based on confusion and ignorance. Confusing absolute advantage with a valid theory of trade sows fear that a nation must protect itself from overwhelming competition. Greatly exaggerated notions of the size of adjustment costs leads to support for protection. Ignorance of the harm done to the world's poor by protection persuades many who support redistribution of income to support protection that harms the majority of those they seek to help. The combination of confusion and ignorance among the 'disinterested' with well-organized special interest groups explains the power of protectionism.

See Also

► Heckscher–Ohlin Trade Theory

Appendix

The general statement of comparative advantage is that on average a country will import goods that are relatively expensive in autarky. Let m denote the vector of excess demands in equilibrium, positive for imports and negative for imports. Let p denote the vector of relative prices in autarky in the home country and let p^* denote the vector of relative prices in autarky in the foreign country. Then the vector inner product $(p - p^*)'m \geq 0$.

The key requirement for the proposition is 'as if' optimization by consumers and producers, leading downward-sloping demand and upward-sloping supply in the generalized sense (the substitution effects matrix of real income-compensated excess demands, m_p^c, is negative semi-definite). If the actual trade equilibrium involves trade distortions, the additional requirement is that trade not be on balance subsidized. Let t be the vector of trade taxes, positive for import taxes and negative for export taxes (and negative for import subsidies and positive for export subsidies). The requirement is $t'm \geq 0$.

The 'buy low, sell high' logic implies that a surplus is captured by trade, so comparative advantage trade is closely linked to the gains from trade. 'As if' optimization means that consumers lower the expenditure required to support given real income by reallocating consumption in trade equilibrium as compared with autarky, while optimization by producers means that income is raised by reallocating production in trade equilibrium as compared with autarky.

Similar comparative advantage statements can be made concerning the factor content of trade; countries tend to import (embodied in goods) the factors that are relatively expensive in autarky (see Neary and Schweinberger 1986).

Bibliography

Anderson, J., and E. van Wincoop. 2003. Gravity with gravitas: A solution to the border puzzle. *American Economic Review* 93: 170–192.

Anderson, J., and E. van Wincoop. 2004. Trade costs. *Journal of Economic Literature* 42: 691–751.

Bernard, A., J. Eaton, B. Jenson, and S. Kortum. 2003. Plants and productivity in international trade. *American Economic Review* 93: 1268–1290.

Bernhofen, D., and J. Brown. 2004. A direct test of the theory of comparative advantage: The case of Japan. *Journal of Political Economy* 112: 48–67.

Davis, D., and D. Weinstein. 2002. An account of global factor trade. *American Economic Review* 91: 1423–1453.

Deardorff, A. 1984. The general validity of the law of comparative advantage. *Journal of Political Economy* 88: 941–957.

Dixit, A., and V. Norman. 1986. Gains from trade without lump-sum compensation. *Journal of International Economics* 21: 111–122.

Eaton, J., and S. Kortum. 2002. Technology, geography and trade. *Econometrica* 70: 1741–1779.

Ethier, W. 1982a. National and international returns to scale in the modern theory of international trade. *American Economic Review* 73: 389–405.

Ethier, W. 1982b. Decreasing costs in international trade and Frank Graham's argument for protection. *Econometrica* 50: 1243–1268.

Feenstra, R. 2003. *Advanced international trade: Theory and evidence*. Princeton: Princeton University Press.

Gresser, E. 2002. Toughest on the poor. *Foreign Affairs* 81(6): 9–14.

Guesnerie, R. 2001. Second best redistributive policies: The case of international trade. *Journal of Public Economic Theory* 3: 15–25.

I

Helpman, E., and P. Krugman. 1985. *Market structure and foreign trade*. Cambridge, MA: MIT Press.
Jones, R. 1987. Heckscher–Ohlin trade theory. In *The New Palgrave: A Dictionary of Economics,* Vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.
Kemp, M., and H. Wan. 1986. Gains from trade with and without lump-sum compensation. *Journal of International Economics* 21: 99–110.
Krugman, P. 1980. Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70: 950–959.
Melitz, M. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71: 1695–1725.
Neary, J., and A. Schweinberger. 1986. Factor content functions and the theory of international trade. *Review of Economic Studies* 53: 421–432.

International Trade, Empirical Approaches to

Stephen J. Redding

Abstract

This article reviews empirical research in international trade, which has undergone a resurgence since the mid-1980s. The article begins with traditional trade empirics, in which cross-country differences in opportunity costs of production (comparative advantage) are the basis for trade, before turning to new trade empirics, in which consumer love of variety and increasing returns to scale give rise to trade in similar goods between similar countries. More recent empirical research has emphasized heterogeneity across products within industries and across individual plants and firms, while other recent work has focused on the political economy of trade policy.

Keywords

Autarky; Bilateral trade; Campaign contributions; Canada–US Free Trade Agreement; Comparative advantage; Constant elasticity of substitution (CES) preferences; Elasticity of substitution; Endowments paradox; Factor content of trade; Factor endowments; Factor price equalization; Factor price equalization theorem; Factor service trade; General Agreement on Tariffs and Trade (GATT); Gravity equation; Heckscher–Ohlin trade theory; Hecksher–Ohlin theorem; International trade; International trade and heterogenous firms; Intra-industry trade; Japan, economics in; Location of production; Missing trade; Most Favored Nation (MFN) principle; Multilateral trade liberalization; Neoclassical trade theory; New trade theory; Non-neutral technology differences; Opportunity costs of production; Outsourcing; Political competition; Protection; Reciprocity; Regional and preferential trade agreements; Ricardo, D.; Rybczynski theorem; Skill-biased technical change; Stolper–Samuelson theorem; Vertical specialization; Wage heterogeneity; World Trade Organization (WTO)

JEL Classifications

F1

Traditional Trade Empirics

The idea that comparative advantage provides an explanation for 'inter-industry trade' (the international exchange of one set of goods for another) dates back to Ricardo (1817), who emphasized technology differences as the source of cross-country variation in opportunity costs of production. While some early empirical studies adopted a Ricardian perspective (for example, MacDougall 1951), much of the empirical analysis of traditional trade frameworks has been concerned with the Heckscher–Ohlin (HO) model (Heckscher 1919; Ohlin 1924). In contrast to its Ricardian counterpart, the HO model assumes that countries have identical technologies, and instead emphasizes variation in country factor endowments and industry factor intensities as the source of differences in opportunity costs of production.

The stylized version of the HO model assumes two factors of production (capital and labour), two

countries (one capital-abundant), and two goods (one capital-intensive at all factor prices). In this stylized case, the model yields four sharp predictions: (*a*) the HO theorem – the capital-abundant country exports the capital-intensive good; (*b*) the factor price equalization theorem – with diversified production, international trade equalizes factor prices; (*c*) the Stolper–Samuelson theorem – with diversified production, an increase in the relative price of the labour-intensive good raises the relative and real return to labour and reduces the relative and real return to capital; (*d*) the Rybczynski theorem – with diversified production, an increase in the endowment of labour leads to a more than proportionate increase in the output of the labour-intensive good and reduces output of the capital-intensive good.

Early empirical examinations of the HO model were loosely motivated by these four theorems. In seeking to test the HO theorem, Leontief (1953) found that US exports were less capital-intensive than US imports, which appeared paradoxical within the confines of the stylized HO model. The key to resolving this paradox in Leamer (1980) was in rigorously deriving the correct empirical predictions directly from the theory. Indeed, a distinguishing feature of recent empirical studies of the HO model has been the derivation of empirical specifications from general equilibrium trade theory and the explicit recognition of the complexity of the model's predictions with many goods and factors of production.

With many goods and factors of production, and in the absence of trade costs, the theorems of the HO model are considerably weaker than in the $2 \times 2 \times 2$ stylized version, and hold only as averages or correlations. We begin by examining predictions for international trade (the generalization of the HO theorem). The many-good, many-factor version of the model does not predict the pattern of trade in individual goods, but does predict the pattern of trade in individual factor services. A country that is abundant in a factor is predicted to be a net exporter of the factor, where factor abundance is defined as an endowment exceeding the country's share of world consumption times the world factor endowment. Therefore, many empirical studies of the HO model have focused on its predictions for net trade in factor services. Following Leamer's (1984) early and influential treatment, Bowen et al. (1987) were the first to observe that a full test of the model's predictions for factor service trade requires three sets of separate data on international trade, factor input requirements and factor endowments. Early empirical results were discouraging from the point of view of the explanatory power of the theory. Bowen et al. (1987) found that the HO model performed no better than a coin toss in predicting the direction of a country's net trade in factor services. In response, Trefler (1993) argued that factor-augmenting technology differences could both explain patterns of trade in factor services and account for cross-country variation in factor prices. Under this hypothesis, first mooted in Leontief (1953), the HO model's predictions for factor service trade and of factor price equalization hold only after one controls for cross-country differences in the efficiency of factors of production. In subsequent work, Trefler (1995) identified two systematic departures between predicted and measured net trade in factor services: (*a*) 'The Case of the Missing Trade', where measured factor services trade is close to zero and much smaller than predicted by the HO model; and (*b*) 'The Endowments Paradox', where rich countries are scarce in most factors and poor countries are abundant in most factors.

One strand of recent research has argued that the HO model's predictions for factor service trade are much closer to the measured values for trade between regions within countries, where the model's assumptions of identical technologies, factor price equalization and identical and homothetic preferences are more likely to be satisfied. Davis et al. (1997) provide evidence supporting the HO model's predictions using data for trade between Japanese regions. A second strand of research has argued that factor-augmenting technology differences are not enough to explain international trade in factor services, but that a reconciliation between theory and data is ultimately possible. Davis and Weinstein (2001) provide evidence that international trade in factor services can be successfully explained if the HO model's assumptions are relaxed to introduce

cross-country differences in technology that vary between industries ('non-neutral' technology differences), trade costs and non-factor price equalization. While predicted and measured net factor service trade has been brought into line, the model is radically transformed by relaxing these assumptions.

We now turn to the predictions of the many-good, many-factor HO model for the international location of production (the generalization of the Rybczynski theorem). With an equal number of goods and factors of production and factor price equalization, the HO model implies a linear relationship between production and factor endowments. Estimating this relationship using cross-country data, Harrigan (1995) finds statistically significant coefficients on factor endowments, but large within-sample prediction errors, suggesting that the model performs poorly in explaining the international location of production. Gandal, Hanson and Slaughter (2004) and Hanson and Slaughter (2002) examine the HO model's prediction that, in an equilibrium where factor prices are pinned down by goods prices, changes in factor endowments should be absorbed through changes in output mix. Using immigration data for Israel and US states, they find some evidence in support of the model's prediction. More recent research reinforces conclusions from the analysis of net factor services trade by suggesting that non-neutral technology differences across industries are important for explaining the international location of production. In an influential paper, Harrigan (1997) estimates an equation for the share of sector in GDP derived from the neoclassical model of trade, which relaxes the assumptions of the HO model to allow for cross-country differences in technology. Both differences in factor endowments and differences in technology that are non-neutral across industries are found to be important in explaining cross-country variation in production structure. Other research finds evidence consistent with multiple cones of diversification within the HO model, where countries or regions specialize in a distinct set of goods, and as a result have different relative factor prices (Schott 2003; Bernard et al. 2005).

We now turn to the relationship between international trade and factor prices, an issue which rose to prominence with the debate about whether the rise in wage inequality in OECD countries since the 1970s is explained by international trade or skill-biased technological change. While the labour economics literature has tended to emphasize the role of skill-biased technological change, the international trade literature has produced mixed findings, as illustrated by the collection of studies in Feenstra (2000). One approach has examined the net factor content of trade and has typically found a relatively minor role for international trade (see, for example, Krugman 2000). Another approach has examined the relationship between relative goods and relative factor prices within the many-good, many-factor version of HO model (the generalization of the Stolper–Samuelson theorem). Here the results have been more sanguine about the contribution of international trade. Leamer (1998) shows how zero-profit conditions and the shares of factors in unit costs for a cross-section of industries can be used to estimate the changes in factor prices mandated by observed changes in goods prices. Assumptions are made about the degree to which improvements in technology are passed through into lower goods prices, and some evidence is found that trade-induced changes in goods prices during the 1970s pushed towards increasing wage inequality in the United States. Feenstra and Hanson (1999) extend the analysis to estimate the contribution of measures of technological change and outsourcing to changes in relative goods and hence through the zero-profit conditions to relative factor prices. In their baseline specification, they estimate that computers explain around 35 per cent of the rise in the relative wages of US non-production workers over the period 1979 to 1990, and outsourcing explains around 15 per cent.

One important difference between international trade and other fields, such as development economics, is that general equilibrium is central to many of the field's theoretical predictions. As it result, it has proved hard to find natural experiments that provide plausible sources of exogenous variation to identify relationships of interest.

Relatedly, many of the predictions of traditional trade theory with many goods and many factors relate to movements from autarky to international trade, but autarky is rarely observed. In two creative papers, Bernhofen and Brown (2004, 2005) exploit the dramatic opening of the Japanese economy in the 19th century from a state of near-complete isolation to test some of the most fundamental predictions of general equilibrium trade theory. In their first paper, they find evidence, supporting the general law of comparative advantage, that an economy's net export vector evaluated at autarky prices is negative. In their second paper, they estimate that, during the final years of Japan's isolation during 1851–3, real income would have had to increase by around eight or nine per cent in order to afford the consumption bundle that the economy could have obtained if it were engaged in international trade during that period.

New Trade Empirics

Although traditional trade theory emphasizes the international exchange of one set of goods for another (inter-industry trade) due to comparative advantage (dissimilar countries), much of international trade involves the two-way exchange of goods within industries (intra-industry trade) between developed nations (similar countries). This apparent disconnect between theory and data was documented in a number of early empirical studies, which examined the extent of intra-industry trade (for example, Grubel and Lloyd 1975) and the volume of trade between similar countries (for example, Linder 1961). This empirical evidence was a key motivation for the 'new trade theory' literature following Krugman (1979, 1980) that explained these features of international trade in terms of consumer love of variety and increasing returns to scale. Firms manufacture differentiated products and concentrate production in a single location, while consumers spread their expenditure across all firms' varieties, giving rise to two-way trade even if countries are identical. Although not the only explanation for intra-industry trade between similar countries (see Davis 1997), the combination of consumer love of variety and increasing returns to scale provided an entirely new intellectual framework for thinking about the causes and consequences of international trade.

In the HO model, the volume of trade is increasing in the extent of dissimilarity in countries' factor endowments, whereas in new trade theory the volume of trade is increasing in the similarity of countries' sizes. Indeed, new trade theory provides rigorous theoretical foundations for the so-called 'gravity equation', in which the volume of trade between two countries is proportional to the product of their sizes and measures of extent of trade frictions. Although the gravity equation had been known for some time to provide an extremely successful empirical explanation for bilateral patterns of international trade (classic early treatments include Tinbergen 1962, and Linnemann 1966), it initially suffered from a lack of theoretical foundations.

New trade theory's prediction that the volume of trade should be proportional to the similarity of country sizes was examined empirically by Helpman (1987) in specifications derived directly from the theory. Using data from 14 OECD countries over the period 1956 to 1981, both bilateral trade and the share of inter-group trade in total trade were found to be strongly increasing in the similarity of country sizes. While this appeared to strongly confirm the predictions of new trade theory, Hummels and Levinsohn (1995) found that the same patterns existed for trade between non-OECD countries, for which new trade theory's assumptions of differentiated products and identical and homothetic preferences appeared less appropriate. One explanation of why the gravity equation appears to work for such diverse groups of countries is that a number of alternative theoretical frameworks, including the HO model, yield this relationship. As argued by Deardorff (1998), the gravity relationship is a basic implication of specialization combined with identical and homothetic preferences. Therefore, the problem is not a lack but rather a surfeit of theoretical foundations. Consistent with this insight, Evenett and Keller (2002) found that increasing returns and factor endowments both

played a role in explaining the empirical success of the gravity equation for a diverse cross-section of developed and developing countries.

The gravity equation has been widely used in empirical work to estimate the impact on trade of a host of frictions, policies and institutions including national borders, transport costs, tariffs, common currencies and the World Trade Organization (WTO). A notable example is McCallum (1995), who finds that trade between Canadian provinces was more than 20 times larger than trade between Canadian provinces and US states, suggesting a surprisingly large impact of national borders on trade. Anderson and Van Wincoop (2002) show, however, that theoretical derivations of the gravity equation imply that bilateral trade depends not only on trade costs between regions themselves ('bilateral resistance') but also on trade costs with all locations ('multilateral resistance'). An implication is that national borders have a larger impact on inter-regional trade than on international trade the smaller a country is and the larger its trade partner. When countries are small, international trade is a large share of overall economic activity. Therefore, the national border has a large effect on multilateral resistance, and so leads to a large reduction in the cost of inter-regional trade relative to international trade. Estimating the gravity equation in a theory-consistent way, Anderson and Van Wincoop (2002) obtain much smaller, though still large, estimates of the trade impact of the Canada–US border.

In the presence of trade costs, an important difference emerges between the predictions of new trade theory and those of traditional trade theory. The combination of consumer love of variety, increasing returns to scale and trade costs in new trade theory generates a 'home market effect', whereby an increase in expenditure leads to a more than proportionate increase in domestic production of a good. The intuition is that increasing returns to scale imply that firms have an incentive to concentrate production, while transport costs imply that they have an incentive to concentrate production close to large markets. In contrast, traditional trade models imply that an increase in expenditure leads at most to a proportionate increase in domestic production if foreign export supply is perfectly inelastic. Otherwise, if the foreign export supply curve is upward sloping, some of the increase in expenditure is satisfied through higher foreign exports and the increase in domestic production is less than proportionate. Using international and Japanese regional data, Davis and Weinstein (1999, 2003) find evidence of home market effects for a number of manufacturing industries, which together account for a substantial share of overall manufacturing activity. Additional evidence in support of home market effects emerges from international trade data in Feenstra et al. (2001) and Hanson and Xiang (2004).

One feature of international trade that appears at first sight hard to reconcile with new trade theory is the large number of zeros between country pairs. The constant elasticity of substitution (CES) preferences and iceberg trade costs in new trade models imply that all country pairs trade a positive quantity of each variety. However, in an analysis of bilateral trade between 161 countries over the period 1970 to 1997, Helpman et al. (2006) find that roughly one half of the country-partner-year observations involve zero trade. A natural explanation for zero bilateral trade flows can be created within new trade theory if firm heterogeneity and fixed trade costs are introduced following Melitz (2003). Depending on the distribution of productivity within countries, firms may or may not find it profitable to incur the fixed costs of exporting to a particular market. Helpman et al. (2006) develop a methodology for estimating the gravity equation that not only controls for multilateral resistance as suggested above, but also controls for the existence of zero bilateral flows and the non-random selection of firms into exporting according to their productivity.

Finally, one key stylized fact about international trade since the Second World War is that it has grown far more rapidly than income. Two potential explanations are reductions in trade barriers following multilateral liberalization or regional integration, and improvements in transportation and communication technologies. Yi (2003) argues that is hard to explain the magnitude of the trade growth using standard trade models, and observed declines in trade barriers

unless one assumes implausibly high elasticities of substitution. However, augmenting standard models to include intermediate inputs enables the growth in trade to be explained with a smaller elasticity of substitution. In the augmented model, tariff reductions decrease the cost of shipping both intermediate inputs and final goods, and so have a magnified impact on overall trade volumes. Indeed, the geographical separation of stages of the production process is one of the distinctive features of trade at the end of the 20th century compared with an earlier era of international integration at the end of the 19th century. This geographical separation of stages of production has been variously referred to as vertical specialization, vertical disintegration, the fragmentation of production, the slicing of the value-added chain, geographical production networks and offshoring. Hummels et al. (2001) define vertical specialization as occurring when the following conditions are satisfied: (*a*) goods are produced in multiple sequential stages; (*b*) two or more countries provide value-added in the good's production sequence; (*c*) at least one country uses imported inputs in its stage of the production process and some of the resulting output is exported. The authors provide empirical evidence of the rapid growth in vertical specialization in the closing decades of the 20th century alongside the rapid growth in overall trade.

The Empirics of Product Trade

The dissemination of highly disaggregated datasets on trade in thousands of individual products (Feenstra et al. 2002; Feenstra et al. 2005) has contributed towards a shift in focus in empirical trade research towards the micro level. For the United States, data are available for over 7,000 seven-digit products of the Tariff Schedule of the United States (TS7) from 1972 to 1988 and for over 10,000 ten-digit products of the Harmonized System (HS10) from 1989 onwards.

In contrast to the empirical research on the HO model discussed above, which emphasizes specialization across products or industries, Schott (2004) provides compelling evidence of specialization *within* products. With US manufacturing imports taken as a whole in 1994, and the unit value ratio (UVR) defined as the ratio of value to quantity, the maximum UVR within products across trade partners is a factor of 24 times greater than the minimum UVR. The UVRs are higher for varieties originating in capital-and skill-abundant countries than for those sourced from labour-abundant countries, consistent with HO-based specialization. Similarly, UVRs are positively associated with the capital intensity of the production techniques that exporters use to produce them. Taken together, these and other findings in the paper suggest that comparative advantage operates at a much finer level of detail than customarily considered.

Another insight that emerges from the product-level trade data is the importance of the 'extensive margin' of the set of goods traded. Hummels and Klenow (2005) decompose variation in countries' aggregate exports into the contributions of the following terms: (*a*) the quantity of each good exported (the 'intensive margin'); (*b*) the set of goods exported (the 'extensive margin'); (*c*) the quality of goods exported. They find that the extensive margin accounts for around 60 per cent of the greater exports of larger economies, while the remaining intensive margin contribution of 40 per cent consists of higher quantities being exported at modestly higher prices. Kehoe and Ruhl (2004) establish an important role for the extensive margin in explaining the growth of trade following trade liberalizations. The set of goods that accounted for only 10 per cent of trade prior to liberalization are found to account for as much as 40 per cent of trade after liberalization. Using micro data from the U.S. Commodity Flow Survey, Hillberry and Hummels (2005) show that trade frictions such as distance reduce the aggregate value of trade primarily through the number of commodities shipped and the number of establishments shipping commodities (the extensive margin) rather than through the average value of shipments (the intensive margin). Together these findings present a number of challenges to standard trade models. For example, in marked contrast to the data, new

trade theory models without firm heterogeneity and fixed costs of trading imply that all of the adjustment to trade frictions occurs through the intensive margin.

Although consumer love of variety is one of the defining features of new trade theory, Broda and Weinstein (2006) were the first to estimate the welfare gains from an increase in the number of varieties imported over time. In their analysis, the product-level trade data is used to measure varieties, defined as the versions of a product supplied by different exporters. The methodology of Feenstra (1994) is extended to estimate separate elasticities of substitution for thousands of products and to evaluate the contribution of new varieties to the US import price index. According to their baseline estimates, conventional price indices that do not correctly control for variety growth overstate the growth in US import prices by around 1.2 percentage points per annum. The estimated contribution to US welfare from an increase in the number of varieties imported over the period 1972 to 2001 is around 2.6 per cent of national income.

The Empirics of Plant and Firm Trade

The analysis of micro data-sets on plants and firms presents additional empirical challenges to traditional and new trade theory, and prompted a wave of subsequent theoretical research. The first set of empirical challenges relates to producer heterogeneity and persistent reallocation. Whereas traditional and new trade theories typically assume a representative firm, micro data-sets reveal vast heterogeneity across plants and firms within narrowly defined industries, in terms of productivity, capital intensity, skill intensity and other characteristics (see for example the survey by Bartelsman and Doms 2000). Similarly, whereas traditional trade theory emphasizes net reallocations of resources between industries in response to exogenous shocks such as trade liberalization, micro data-sets reveal persistent job creation and job destruction in all industries even in the apparent absence of exogenous shocks. Additionally, job creation and job destruction are positively correlated across industries, implying that rates of gross job creation and destruction are large relative to the net reallocation emphasized in traditional trade theory. An implication of these findings is that the changes in employment across plants and firms are greater than those required to achieve the observed between-industry reallocation of resources ('excess job reallocation'), implying substantial within-industry reallocations of resources (see in particular Davis et al. 1998).

The second set of empirical challenges relates to the export behaviour of plants and firms. Traditional trade theory predicts net exports in one set of industries and net imports in another set of industries. New trade theory implies that all firms export as a result of consumer love of variety and increasing returns to scale. Yet, in micro data-sets, all manufacturing industries display a mix of exporters and non-exporters (see Bernard and Jensen 1995). Moreover, exporters are systematically more productive, more capital-intensive and more skill-intensive than non-exporters (see again Bernard and Jensen 1995). These findings have led to considerable debate as to whether high-performing firms become exporters or whether exporting leads to improved firm performance. The current consensus favours causality running from good firm performance to exporting (selection into export markets): see, for example, Bernard and Jensen (1999), Clerides et al. (1998) and Roberts and Tybout (1997).

The third set of empirical challenges relates to evidence from trade liberalizations in both developed and developing countries. Despite traditional trade theory's emphasis on between-industry reallocations of resources, one of the central findings from empirical studies of trade liberalizations is the importance of within-industry reallocations of resources across plants and firms. In an influential paper, Pavcnik (2002) finds that between-plant reallocations of resources account for around two-thirds (12.4 percentage points) of the 19 per cent increase in aggregate productivity in the Chilean manufacturing sector following the trade liberalization of the late 1970s and early 1980s. Similarly, Trefler (2004) finds an important role for reallocation in accounting for

the improvement in aggregate productivity in Canadian manufacturing in the aftermath of the Canada–US free trade agreement.

Together these empirical challenges have led to the development of new theoretical frameworks incorporating firm heterogeneity into both traditional and new trade theory (see in particular Bernard et al. 2003; Melitz 2003). The interplay between the econometric analysis of micro datasets on plants and firms and the theoretical analysis of firm-based responses to international trade is one of the most exciting areas of ongoing research.

The Empirics of Trade Policy

A final area of rapid recent progress is the empirical analysis of trade policy. A number of alternative approaches to modelling the political economy of trade policy have been taken, including median-voter theories, models where the government trades off political support from industry against consumer dissatisfaction, theories of lobbying by special interest groups, and models of electoral contribution.

One of the most influential lines of research follows the seminal work of Grossman and Helpman (1994). In their model, campaign contributions are designed to influence policy choices. Interest groups move first and offer politicians campaign contributions that depend on their policy stance. Politicians next maximize a political objective function which depends on both campaign contributions and social welfare. The political objective function is derived from microeconomic foundations within a model of electoral competition. The model yields a structural equation in which the level of protection depends on the political organization of the industry, the ratio of domestic output in the industry to net trade, and the elasticity of import demand or export supply. Goldberg and Maggi (1999) estimate the structural relationship implied by the Grossman and Helpman model and find broad empirical support. In particular, the pattern of protection differs markedly between politically organized and non-organized industries, though the implied weight on social welfare relative to political contributions is larger than expected. One of the distinctive features of recent empirical work in this area is again the rigorous derivation of empirical specifications from theoretical predictions. Gawande and Krishna (2003) survey both the recent empirical evidence and the results of earlier and more ad hoc empirical specifications.

A related theoretical literature has sought to model the politics of international trade agreements (for example, Grossman and Helpman 1995; Krishna 1998; McLaren 2002). One issue that has attracted particular attention is the extent to which regional preferential trade agreements reinforce or retard multilateral trade liberalization. Theoretical research has also examined the microeconomic foundations for observed features of international trade institutions such as the General Agreement on Tariffs and Trade (GATT) and the World Trade Organization (WTO) (see in particular Bagwell and Staiger 1999, 2001). Two key features are reciprocity and non-discrimination (the Most Favored Nation, MFN, principle). Empirical work in this area remains in its infancy and offers an exciting prospect for the future. In an analysis of US trade policy, Limao (2006) finds evidence that preferential trading blocs have acted as stumbling blocks for multilateral liberalization.

See Also

- Factor Content of Trade
- Heckscher–Ohlin Trade Theory
- International Trade and Heterogeneous Firms
- International Trade Theory
- Ricardian Trade Theory
- Trade Policy, Political Economy of

Bibliography

Anderson, J., and E. van Wincoop. 2002. Gravity with gravitas: A solution to the border puzzle. *American Economic Review* 93: 170–192.

Bagwell, K., and R. Staiger. 1999. An economic theory of GATT. *American Economic Review* 89: 215–248.

Bagwell, K., and R. Staiger. 2001. Domestic policies, national sovereignty and international economic institutions. *Quarterly Journal of Economics* 116: 519–562.

Bartelsman, E., and M. Doms. 2000. Understanding productivity: Lessons from longitudinal microdata. *Journal of Economic Literature* 38: 569–594.

Bernard, A.B.., J. Eaton, and S.S. Kortum. 2003. Plants and productivity in international trade. *American Economic Review* 93: 1268–1290.

Bernard, A.B.., and J.B. Jensen. 1995. Exporters, jobs, and wages in US manufacturing: 1976–87. *Brookings Papers on Economic Activity: Microeconomics* 67–112.

Bernard, A.B.., and J.B. Jensen. 1999. Exceptional exporter performance: cause, effect, or both? *Journal of International Economics* 47: 1–25.

Bernard, A.B.., S.J. Redding, and P.K. Schott. 2005. Factor price equality and the economies of the United States. Discussion Paper No. 5111. London: CEPR.

Bernhofen, D., and J. Brown. 2004. A direct test of the theory of comparative advantage: The case of Japan. *Journal of Political Economy* 112: 48–67.

Bernhofen, D., and J. Brown. 2005. An empirical assessment of the comparative advantage gains from trade: Evidence from Japan. *American Economic Review* 95: 208–225.

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

Broda, C., and D.E. Weinstein. 2006. Globalization and the gains from variety. *Quarterly Journal of Economics* 121: 541–586.

Clerides, S., S. Lach, and J. Tybout. 1998. Is learning by exporting important? Micro-dynamic evidence from Columbia, Mexico and Morocco. *Quarterly Journal of Economics* 113: 903–947.

Davis, D. 1997. Critical evidence on comparative advantage? North–north trade in a multilateral world. *Journal of Political Economy* 105: 1051–1060.

Davis, S., J. Haltiwanger, and S. Schuh. 1998. *Job creation and destruction*. Cambridge, MA: MIT Press.

Davis, D., and D. Weinstein. 1999. Economic geography and regional production structure: An empirical investigation. *European Economic Review* 43: 379–407.

Davis, D., and D. Weinstein. 2001. An account of global factor trade. *American Economic Review* 91: 1423–1453.

Davis, D., and D. Weinstein. 2003. Market access, economic geography, and comparative advantage: An empirical assessment. *Journal of International Economics* 59: 1–23.

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

Deardorff, A.V. 1998. Determinants of bilateral trade: does gravity work in a neoclassical world? In *The regionalization of the world economy*, ed. J. Frankel. Chicago: NBER and Chicago University Press.

Evenett, S., and W. Keller. 2002. On theories explaining the success of the gravity equation. *Journal of Political Economy* 110: 281–316.

Feenstra, R.C. 1994. New product varieties and the measurement of international prices. *American Economic Review* 84: 157–177.

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

Feenstra, R., and G. Hanson. 1999. The impact of outsourcing and high-technology capital on wages: Estimates for the United States, 1979–80. *Quarterly Journal of Economics* 114: 907–940.

Feenstra, R., J. Markusen, and A. Rose. 2001. Understanding the home market effect and the gravity equation. *Canadian Journal of Economics* 34: 430–447.

Feenstra, R.C., J. Romalis, and P.K. Schott. 2002. U.S. imports, exports, and tariff data, 1989–2001. Working Paper No. 9387. Cambridge, MA: NBER.

Feenstra, R.C., R.E. Lipsey, H. Deng, A.C. Ma, and H. Mo. 2005. World trade flows: 1962–2000. Working Paper No. 11040. Cambridge, MA: NBER.

Gandal, N., G. Hanson, and M. Slaughter. 2004. Technology, trade, and adjustment to immigration in Israel. *European Economic Review* 48: 403–428.

Gawande, K., and P. Krishna. 2003. The political economy of trade policy: Empirical approaches. In *Handbook of international trade*, ed. E.K. Choi and J. Harrigan. Oxford: Basil Blackwell.

Goldberg, P., and G. Maggi. 1999. Protection for sale: An empirical investigation. *American Economic Review* 89: 1135–1155.

Grossman, G.M., and E. Helpman. 1994. Protection for sale. *American Economic Review* 84: 833–850.

Grossman, G.M., and E. Helpman. 1995. The politics of free trade agreements. *American Economic Review* 85: 667–690.

Grubel, H., and P. Lloyd. 1975. *Intra-industry trade: The theory and measurement of international trade in differentiated products*. London: Macmillan.

Hanson, G., and M. Slaughter. 2002. Labor market adjustment in open economies: Evidence from U.S. states. *Journal of International Economics* 57: 3–29.

Hanson, G., and C. Xiang. 2004. The home market effect and bilateral trade patterns. *American Economic Review* 94: 1109–1129.

Harrigan, J. 1995. Factor endowments and the international location of production: Econometric evidence for the OECD, 1970–85. *Journal of International Economics* 39: 123–141.

Harrigan, J. 1997. Technology, factor supplies, and international specialisation: Estimating the neoclassical model. *American Economic Review* 87: 475–494.

Heckscher, E.F. 1919. The effect of foreign trade on the distribution of income. *Economisk Tidsckrift*. In *Heckscher–Ohlin trade theory*, ed. E.F. Heckscher, and B. Ohlin. Cambridge, MA: MIT Press, 1991.

Helpman, E. 1987. Imperfect competition and international trade: Evidence from fourteen industrial countries. *Journal of the Japanese and International Economies* 1: 62–81.

Helpman, E., M. Melitz, and Y. Rubinstein. 2006. *Trading partners and trading volumes*. Mimeo: Harvard University.
Hillberry, R.. and D. Hummels. 2005. Trade responses to geographic frictions: a decomposition using micro-data. Working Paper No. 11339. Cambridge, MA: NBER.
Hummels, D., J. Ishii, and K.-M. Yi. 2001. The nature and growth of vertical specialization in world trade. *Journal of International Economics* 54: 75–96.
Hummels, D., and P. Klenow. 2005. The variety and quality of a nation's exports. *American Economic Review* 95: 704–723.
Hummels, D., and J. Levinsohn. 1995. Monopolistic competition and international trade: Reconsidering the evidence. *Quarterly Journal of Economics* 110: 799–836.
Kehoe, T., and K. Ruhl. 2004. How important is the new goods margin in international trade? Working Paper, Federal Reserve Bank of Minneapolis.
Krishna, P. 1998. Regionalism and multilateralism: A political economy approach. *Quarterly Journal of Economics* 113: 227–251.
Krugman, P.R. 1979. Increasing returns, monopolistic competition, and international trade. *Journal of International Economics* 9: 469–479.
Krugman, P.R. 1980. Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70: 950–959.
Krugman, P.R. 2000. Technology, trade and factor prices. *Journal of International Economics* 50: 51–72.
Leamer, E. 1980. The Leontief paradox, reconsidered. *Journal of Political Economy* 88: 495–503.
Leamer, E. 1984. *Sources of international comparative advantage*. Cambridge, MA: MIT Press.
Leamer, E. 1998. In search of Stolper–Samuelson linkages between international trade and lower wages. In *Imports, exports, and the American worker*, ed. S. Collins. Washington, DC: Brookings Institution Press.
Leontief, W. 1953. Domestic production and foreign trade: The American capital position re-examined. *Proceedings of the American Philosophical Society* 97: 332–349.
Limao, N. 2006. Preferential trade agreements as stumbling blocks for multilateral trade liberalization: Evidence for the United States. *American Economic Review* 96: 896–914.
Linder, S.B. 1961. *An essay on trade and transformation*. New York: Wiley.
Linnemann, H. 1966. *An econometric study of international trade flows*. Amsterdam: North-Holland.
McCallum, J. 1995. National borders matter: Canada–US regional trade patterns. *American Economic Review* 85: 615–623.
MacDougall, G.D.A. 1951. British and American exports: A study suggested by the theory of comparative costs. *Economic Journal* 61: 697–724.
McLaren, J. 2002. A theory of insidious regionalism. *Quarterly Journal of Economics* 117: 571–608.
Melitz, M.J. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71: 1695–1725.
Ohlin, B. 1924. The theory of trade. In *Heckscher–Ohlin trade theory*, ed. E.F. Heckscher, and B. Ohlin. Cambridge, MA: MIT Press, 1991.
Pavcnik, N. 2002. Trade liberalization, exit, and productivity improvement: Evidence from Chilean plants. *Review of Economic Studies* 69: 245–276.
Ricardo, D. 1817. *The principles of political economy and taxation*. London: John Murray.
Roberts, M.J., and J. Tybout. 1997. The decision to export in Colombia: An empirical model of entry with sunk costs. *American Economic Review* 87: 545–564.
Schott, P.K. 2003. One size fits all? Heckscher–Ohlin specialization in global production. *American Economic Review* 93: 686–708.
Schott, P.K. 2004. Across-product versus within-product specialization in international trade. *Quarterly Journal of Economics* 119: 647–678.
Tinbergen, J. 1962. *Shaping the world economy: Suggestions for an international economic policy*. New York: Twentieth Century Fund.
Trefler, D. 1993. International factor price differences: Leontief was right! *Journal of Political Economy* 101: 961–987.
Trefler, D. 1995. The case of the missing trade and other mysteries. *American Economic Review* 85: 1029–1046.
Trefler, D. 2004. The long and short of the Canada–U.S. Free trade agreement. *American Economic Review* 94: 870–895.
Yi, K. 2003. Can vertical specialization explain the growth of world trade? *Journal of Political Economy* 111: 52–102.

Internet and the Offline World

Avi Goldfarb

Abstract

This article emphasises that a key to understanding the (net) benefits of the Internet is to remember that all online activity has an offline context. People live their lives offline. Therefore, the fall in communication costs and the fall in distribution costs associated with the diffusion of the Internet had a heterogeneous impact across locations.

Keywords

Agglomeration; Communication costs; Distribution costs; E-commerce; Internet; Online–offline; Regulation

JEL Classifications

L81; L86; R11

Introduction

There is a large and growing literature in economics and management that studies the benefits and costs of the Internet to firms, consumers, academics, and workers. Researchers have studied online purchasing habits, online advertising, online communication, online work, and online business strategy; and this research has built our understanding of how people behave online. Importantly, however, much of it has focused on data and settings that include only online activities.

In this article, I summarise a literature that argues that a key to understanding the (net) benefits of the Internet is to remember that all online activity has an offline context. People live their lives offline and therefore an understanding of offline opportunities, desires, and positions provides a more complete picture of the impact of the online communications channel. (Related literature reviews include Scott Morton (2006) and Lieber and Syverson (2012)).

Cairncross (1997) was among the first to recognise how the Internet helped substitute for several offline activities, emphasising how the Internet would bring about the 'Death of Distance'. This argument to some degree echoed McLuhan's (1962) much earlier writing about the impact of electronic media (mainly television) on global culture. Friedman's (2005) discussion of the 'Flat World' also is motivated by the reduced importance of the local environment due to advances in information and communication technology.

Much of the academic literature on this topic has emphasised the heterogeneous effects of Internet technology across locations. The next two sections discuss how reductions in communications costs and distribution costs have such heterogeneous effects across locations. For example, digital distribution of movies has a much larger impact on the types of movies available to consumers in North Dakota than in Manhattan. The next section discusses how different local regulatory environments mean that the impact of the technology varies across locations. The penultimate section emphasises that even if the benefits from Internet technology were uniform across locations, the costs of adopting Internet technology (for businesses and consumers) are lowest in large urban areas. The article concludes by reemphasising the importance of heterogeneous offline environments to understanding the impact of Internet technology.

Communication Costs

The Internet is a communications technology. Specifically, it is a series of protocols that allow computers to communicate with each other. These protocols, in turn, enable people to communicate in new, often relatively inexpensive, ways. For example, the marginal cost of sending an email is effectively zero. In contrast, the marginal cost of sending a letter is much higher. Furthermore, while the cost of sending a letter increases as borders are crossed, and often as distance increases, the cost of sending an email does not change with distance. Thus Internet technology has meant a lowering of communications costs, whether the communication was local or distant.

However, the impact of this reduction in communications costs differs by location and by type of communication. In particular, in order to understand the impact of the reduction in communications costs, researchers have focused on relative impact across locations. While communications costs fall everywhere, the size of the reduction depends on the alternative communications technologies available.

One possibility is that the Internet substitutes for things that are available offline. Online

advertising can substitute for offline advertising, online news can substitute for offline news, and online communication can substitute for offline communication. For example, Goldfarb and Tucker (2011a, b) show that online advertising is both more expensive and more effective in locations where regulations make offline advertising difficult; Sinai and Waldfogel (2004) show that online news substitutes for offline news; and Forman and van Zeebroeck (2012) demonstrate the online communication can overcome barriers to offline communication. Even when there are barriers to using the Internet to substitute for an offline activity, new markets can arise to overcome these barriers. For example, Jin and Kato (2007) use evidence from the sports memorabilia industry to demonstrate that asymmetric information issues in the online market can be overcome by the establishment of new markets for information, thereby enabling the online market to provide a reasonable substitute for the offline market.

An alternative to substitution is that the Internet complements things that are available locally. Goldmanis et al. (2010) emphasise that the Internet reduces search costs, benefiting the highest quality offline businesses while hurting others. Thus the Internet complements high-quality local businesses.

Blum and Goldfarb (2006) demonstrate that Internet surfing is disproportionately local, likely because consumer tastes are spatially correlated. While the paper focuses on the foreign (international) websites visited by consumers, the idea is best summarised by recognising that the consumers most likely to visit web pages associated with the Boston Red Sox live disproportionately close to Boston. One implication is that websites may arise locally to serve the tastes of nearby consumers. Similarly, in examining reviews on Amazon, Forman et al. (2008) show that people trust local reviews most, again suggesting spatially correlated tastes.

Hampton and Wellman (2002) argue that social networks are disproportionately local. Thus, Internet technology led the relative price of long distance communication to fall relative to local communication, but Internet communication remains local. Email is often sent within offices and households and Facebook social networks are highly local.

Gaspar and Glaeser (1998) recognised early the potential for the Internet to either substitute for or complement local activities. They pointed out that the impact of such a fall in communications costs does not necessarily increase distant communication relative to local communication, and the impact depends on the particular characteristics of the situation. Glaeser and Ponzetto (2007) argue that complementarities may be more important to driving overall economic activity than substitutes, and Forman et al. (2012) provides some evidence in support of their hypothesis.

This tension between the potential for the Internet to substitute for and complement local activities may be best represented by two papers that examined the impact of electronic communication on collaboration between researchers. Forman and van Zeebroeck (2012) examine the impact of Internet technology on collaboration within a large company. Specifically, they combine information on the location of patent holders within a company with information on the adoption of Internet technology (in the 1990s) across the company's geographically dispersed establishments. Their evidence suggests that the Internet increased collaboration within the company, especially between distant establishments. Agrawal and Goldfarb (2008) examine the impact of an early type of Internet technology, Bitnet, on collaboration between academic researchers. Their evidence also suggests that electronic communication increased collaboration; however, the effect is strongest between co-located universities. Thus, in contrast to Forman and van Zeebroeck's results within a firm, the results across universities suggest that co-location matters. While there are several possible interpretations of the different conclusions of Agrawal and Goldfarb (2008) and Forman and van Zeebroeck (2012), the most likely explanations emphasise differences across the empirical settings. In particular, while both examine the impact of a reduction in communications costs on research collaboration, the marginal benefit to communicating over distances within a firm may exceed the benefits across universities.

For example, within a firm, the availability of local substitutes might be more limited.

Furthermore, the reduction in communications cost facilitated by Internet technology has enabled the development of platforms that reduce frictions that can inhibit transactions. These platforms can have a different impact depending on the availability of local substitutes and complements across locations. For example, online auction platforms provide a communication (and transaction) platform for buyers and sellers to meet. Hortaçsu et al. (2009) examine the geographic patterns of trade on two large online auction platforms: eBay and MercadoLibre. While they show that the distance between buyer and seller is an important deterrent to trade on these platforms, these effects are smaller than those observed in studies of offline transactions. Thus, the reduction in communication costs facilitated by Internet technology disproportionately enables distant transactions relative to the pre-existing geographic distribution of transactions. The online setting enables transactions that might otherwise not be feasible, though distance does still play a role.

Crowdfunding platforms, such as Kickstarter and Sellaband, are another type of Internet-enabled platform. These websites enable entrepreneurs and artists to match with investors in settings that facilitate small scale investments in very early stage projects. Similar to the findings in Hortaçsu et al. (2009), Agrawal et al (2012) document substantial geographic distance between artists and investors on the Sellaband music crowdfunding platform. Importantly, however, this does not mean that geographic distance is irrelevant on crowdfunding platforms. Agrawal et al. (2012) show that the first investors in an artist are typically co-located with the artists. It is only once the artist reaches prominence on the website (partly due to the earlier local investors) that distant investments accelerate. Agrawal et al. explain this difference between local and distant investors with evidence on the offline social networks of the artists: the artists tend to know the early investors personally and because social networks are disproportionately local, early investors are disproportionately local.

This section has discussed findings that the reduction in communication costs associated with the Internet had heterogeneous effects on interactions. Broadly, while it increased both local and distant interaction, the relative impact varied across empirical settings.

Distribution Costs

Internet technology can reduce distribution costs. First, digital distribution means that the cost of sending digitised goods to customers (such as music, news and movies) is close to zero. Second, complementarities between digital ordering and courier systems mean that the cost of sending some physical goods to customers can be lower on online platforms than offline platforms.

Balasubramanian's (1998) model of competition between direct sales channels and offline retail stores provides a useful way to think about the impact of the Internet on retailing. In his model, online (direct) retailers are equally attractive to all customers. In contrast, offline retailers are most attractive to people who live near them. This means that, even though online retailing is equally attractive to all customers, the competition from offline stores is fiercest for customers that live near offline stores. Therefore customers that live near offline stores buy from those stores and customers that do not live near offline stores buy online. Forman et al. (2009) provide empirical evidence supporting this model, demonstrating that offline store openings change the distribution of sales on Amazon away from the top sellers. This implies that while the online channel means that total (firm and customer) distribution costs fall for all locations, the impact of the arrival of the online channel will be largest for those who live far from major retail centers. Brynjolfsson et al. (2009) similarly show substitution between online and offline channels in the apparent industry. Choi and Bell (2011) also demonstrate substitution between online and offline purchases in the diaper category. More importantly, they show that online retail has the ability to provide niche services in places without rich variety in a specific category. In particular, they

show that the best online customers for specialty diapers live in places with few babies. The ability of the online channel to substitute for the offline channel is therefore particularly acute for products where distribution varies with local demand.

For digital products that do not need to be shipped to be consumed, the Internet reduces distribution costs to near zero. Blum and Goldfarb (2006) show that for such products, the effects of borders and distance may disappear. In particular, Blum and Goldfarb show that for digital products that are not taste-based, distance does not affect web browsing behaviour. Thus, low distribution costs do eliminate the role of distance. That said, as mentioned above, this is not true of taste-based products such as music or movies. For those products, distance matters because tastes are spatially correlated.

Regulation

Regulation matters to consumer and business choices. Importantly, regulations vary across countries, states, and even cities. These regulations change the marginal benefit to online activities relative to offline activities. The earliest and perhaps most prominent stream of this literature examines the role of local sales taxes on online purchasing. Goolsbee (2000) demonstrated that the tax-free status of most online purchases played an important role in the rise of electronic commerce. In doing so, he provided evidence that the places with the most online purchases were places with high local sales taxes. Anderson et al. (2010) provided evidence that search processes play an important role in explaining the high rate of online purchases in states with sales taxes. Ellison and Ellison (2009) also find high taxes to be an important motivator for online purchases.

Aside from taxes, advertising and privacy regulations affect the use of the Internet in other ways. Goldfarb and Tucker (2011a) show that regulations banning personal injury lawyers from contacting customers directly appear to raise the price of search engine advertising. In other words, the online channel appears to substitute for the offline channel, partially circumventing the regulation. In a different context, Goldfarb and Tucker (2011b) demonstrate that regulations banning alcohol companies from using billboards appear to increase the effectiveness of online display advertising. Again, the online channel overcomes the barriers to advertising presented by the regulation. Privacy regulations also impact online behaviour. Goldfarb and Tucker (2011c) demonstrate that strict privacy regulations in the European Union appeared to reduce the effectiveness of online advertising.

Costs of Internet Adoption

This last, and briefest, section summarises research that emphasises that the cost of adopting the Internet (to consumers and businesses) varies across locations. Greenstein (2000) demonstrated that competition between Internet Service Providers is highly variable around the USA. Some locations had fierce competition while others had very little competition. Recent research has shown that this result persisted for many years (Mack and Grubesic 2009; Grubesic 2012; Prieger and Hu 2008). This variation in competition appears to lead to different prices for Internet access across locations (Greenstein 2000). For businesses, the main costs of adoption of Internet services involve training workers to use the technology efficiently. Forman et al. (2005, 2008) demonstrated that these costs are heterogeneous across locations within the USA. In particular, cities have much lower costs of adoption than rural areas. Overall, for both consumers and firms, costs of Internet adoption are lowest in urban areas and can be much higher in suburban and especially rural areas.

Summary and Conclusion

This article has emphasised the important role of an individual's offline surroundings in understanding their online behaviour. While the Internet can reduce communications costs and distribution costs, the impact of this reduction is heterogeneous across locations. The Internet can substitute for offline

activities, or complement them. Regulation affects how online and offline interact, and the costs of the adopting the technology itself vary across locations.

Looking forward, there are many reasons to expect more of the same: that heterogeneity in the offline environment will continue to impact online behaviour. With the rise of the mobile Internet, the role of offline surroundings may become even more important to understanding online behaviour (Ghose et al. 2012). Generally, without understanding an individual's offline opportunities, needs and desires, it is difficult to comprehend their online activities.

See Also

- Computer Industry
- Electronic Commerce
- Gravity Models
- Information Technology and the World Economy
- Internet, Economics of The
- Regional Development, Geography of
- Skill-Biased Technical Change
- Urban Agglomeration

Bibliography

Agrawal, A., and A. Goldfarb. 2008. Restructuring research: Communication costs and the democratization of university innovation. *American Economic Review* 98(4): 1578–1590.

Agrawal, A., C. Catalini, and A. Goldfarb. 2012. *Crowdfunding: Social frictions in the flat world?* Working paper, University of Toronto.

Anderson, E., N. Fong, D. Simester, and C. Tucker. 2010. How sales taxes affect customer and firm behavior: The role of search on the Internet. *Journal of Marketing Research* 47(2): 229–239.

Balasubramanian, S. 1998. Mail versus mall: A strategic analysis of competition between direct marketers and conventional retailers. *Marketing Science* 17(3): 181–195.

Blum, B., and A. Goldfarb. 2006. Does the Internet defy the law of gravity? *Journal of International Economics* 70(2): 384–405.

Brynjolfsson, E., Y.J. Hu, and M. Rahman. 2009. Battle of retail channels: How product selection and geography drive cross-channel competition. *Management Science* 55(11): 1755–1765.

Cairncross, F. 1997. *The death of distance*. Cambridge, MA: Harvard University Press.

Choi, J., and D. Bell. 2011. Preference minorities and the Internet. *Journal of Marketing Research* 48(4): 670–682.

Ellison, G., and S.F. Ellison. 2009. Tax sensitivity and home state preferences in Internet purchasing. *American Economic Journal: Economic Policy* 1(2): 53–71.

Forman, C., and N. van Zeebroeck. 2012. From wires to partners: How the internet has Fostered R&D collaborations within firms. *Management Science* 58(8): 1549–1568.

Forman, C., A. Goldfarb, and S. Greenstein. 2005. How did location affect adoption of the commercial Internet: Global village vs. urban leadership. *Journal of Urban Economics* 58(3): 389–420.

Forman, C., A. Goldfarb, and S. Greenstein. 2008a. Understanding the inputs into innovation: Do cities substitute for internal firm resources? *Journal of Economics and Management Strategy* 17(2): 295–316.

Forman, C., A. Ghose, and B. Wiesenfeld. 2008b. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* 19(3): 291–313.

Forman, C., A. Ghose, and A. Goldfarb. 2009. Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management Science* 54(1): 47–57.

Forman, C., A. Goldfarb, and S. Greenstein. 2012. The Internet and local wages: A puzzle. *American Economic Review* 102(1): 556–575.

Friedman, T. 2005. *The world is flat: A brief history of the twenty-first century*. New York: Farrar, Straus, and Giroux.

Gaspar, J., and E. Glaeser. 1998. Information technology and the future of cities. *Journal of Urban Economics* 43: 136–156.

Ghose, A., A. Goldfarb, and S.P. Han. 2012. How is the mobile internet different? Search costs and local activities. *Information Systems Research*, in press.

Glaeser, E. L., and Ponzetto, G. A. M. 2007. *Did the death of distance hurt detroit and help New York?* NBER Working Paper #13710.

Goldfarb, A., and C. Tucker. 2011a. Advertising bans and the substitutability of online and offline advertising. *Journal of Marketing Research* 48(2): 207–228.

Goldfarb, A., and C. Tucker. 2011b. Search engine advertising: Channel substitution when pricing ads to context. *Management Science* 57(3): 458–470.

Goldfarb, A., and C. Tucker. 2011c. Privacy regulation and online advertising. *Management Science* 57(1): 57–71.

Goldmanis, M., A. Hortacsu, C. Syverson, and O. Emre. 2010. E-commerce and the market structure of retail industries. *Economic Journal* 120(545): 651–682.

Goolsbee, A. 2000. In a world without borders: The impact of taxes on internet commerce. *The Quarterly Journal of Economics* 115(2): 561–576.

Greenstein, S. 2000. Building and delivering the virtual world: Commercializing services for internet access. *Journal of Industrial Economics* 48(4): 391–411.

Grubesic, T.H. 2012. The national broadband map: Data limitations and implications for public policy evaluation. *Telecommunications Policy* 36: 113–126.
Hampton, K., and B. Wellman. 2002. Neighboring in Netville: How the Internet supports community and social capital in a wired suburb. *City and Community* 2(3): 277–311.
Hortaçsu, A., F. Asís Martínez-Jerez, and J. Douglas. 2009. The geography of trade in online transactions: Evidence from eBay and MercadoLibre. *American Economic Journal: Microeconomics* 1(1): 53–74.
Jin, G.Z., and A. Kato. 2007. Dividing online and offline: A case study. *Review of Economic Studies* 74(3): 981–1004.
Lieber, E., and C. Syverson. 2012. Online v. offline competition. In *Oxford handbook of the digital economy*, ed. M. Peitz and J. Waldfogel. Oxford: Oxford University Press.
Mack, E.A., and T.H. Grubesic. 2009. Forecasting broadband provision. *Information Economics and Policy* 21(4): 297–311.
McLuhan, M. 1962. *The Gutenberg galaxy*. Toronto: University of Toronto Press.
Prieger, J., and W.-M. Hu. 2008. The broadband digital divide and the nexus of race, competition, and quality. *Information Economics and Policy* 20(2): 150–167.
Scott Morton, F. 2006. Consumer benefit from use of the Internet. In *Innovation policy and the economy*, ed. A.B.. Jaffe, J. Lerner, and S. Stern, Vol. 6, 67–90. Cambridge, MA: MIT Press.
Sinai, T., and J. Waldfogel. 2004. Geography and the Internet: Is the Internet a substitute or a complement for cities? *Journal of Urban Economics* 56: 1–24.

Internet, Economics of the

Nicholas Economides

Abstract

We discuss salient economic aspects of the Internet, including the possible abolition of net neutrality by local broadband access networks as well as potential incompatibilities and degradation of connectivity in the Internet backbone.

Keywords

Antitrust; Cartels; Common standards; Innovation; Interconnection; Internet, economics of the; Market power; Net neutrality; Network effects; Network externalities; Price discrimination; Proprietary standards; Switching costs; Two-sided pricing; Universal connectivity

JEL Classifications

D85

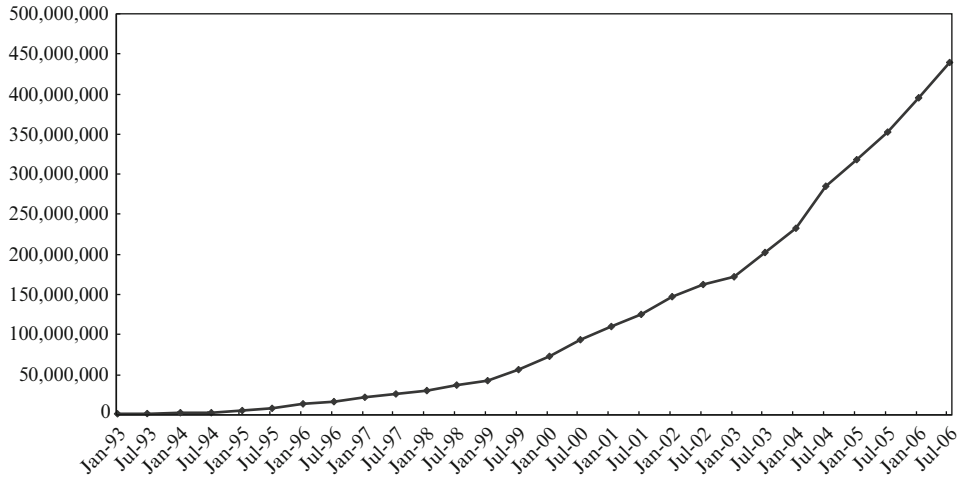
The Internet is a global network of interconnected networks that connect computers. The Internet allows data transfers as well as the provision of a variety of interactive real-time and time-delayed telecommunications services. Internet communication is based on common and public protocols. Hundreds of millions of computers are presently connected to the Internet. Figure 1 shows the expansion of the number of computers connected to the Internet.

The vast majority of computers owned by individuals or businesses connect to the Internet through commercial 'Internet service providers (ISPs). Educational institutions and government departments are also connected to the Internet but typically do not offer commercial ISP services. Users connect to the Internet either by dialing their ISP, connecting through cable modems, or residential 'digital subscriber line' (DSL), or through corporate networks. Typically, routers and switches owned by the ISP send the caller's packets to a local 'point of presence' (POP) of the Internet. Dial-up, cable modem, and DSL access POPs as well as corporate networks dedicated access circuits connect to high-speed hubs. High-speed circuits, leased from or owned by telephone companies, connect the high speed hubs forming an 'Internet backbone network'.

The Internet is based on three basic separate levels of functions of the network:

- the hardware/electronics level of the physical network;
- the (logical) network level where basic communication and interoperability is established; and
- the applications/services level.

Thus, the Internet separates the network interoperability level from the applications/services



Internet, Economics of the, Fig. 1 Internet survey host count, 1993–2006 (*Source*: Internet Systems Consortium). (Online) Available at http://www.isc.org. Accessed 29 January 2007

level. Unlike earlier centralized digital electronic communications networks, such as CompuServe, AT&T Mail, Prodigy, and early America On Line (AOL), the Internet allows a large variety of applications and services to be run 'at the edge' of the network and not centrally.

Residential Broadband Access Networks and Net Neutrality

Users pay ISPs for access to the whole Internet. Similarly, ISPs pay backbones for access to the whole Internet. ISPs pay per month for a pipe of a certain bandwidth, presumably according to their expected use. When digital content, for example, is downloaded by consumer A from provider B, both sides, that is, both A and B, pay. Consumer A pays to his ISP through his monthly subscription, and provider B pays similarly. In turn, ISPs pay to their respective backbones through their monthly subscriptions. The present regime on the Internet does not distinguish in terms of price (or in any other way) between bits or information packets depending on the services that these bits and packets are used for. This regime, called 'net neutrality', has prevailed on the Internet since its inception. Presently, a bit or information packet used for 'voice over Internet protocol' (VOIP), for search, email, for an image or for a video is priced equally as a part of the large number of packets that correspond to the subscription services of the originating and terminating ISP.

Taking advantage of a change in regulatory rules by the Federal Communications Commission that reclassified the Internet as an 'information service' rather than a 'telecommunications service', AT&T, Verizon and cable TV networks advocate price discrimination based on which application and on which provider the bits they transport come from. These local broadband access networks would like to abolish the regime of non-discrimination which has been called 'net neutrality' and substitute for it a complex price discrimination schedule where, besides the basic service for transmission of bits, there will be additional charges by the Internet access network levied to the originating party (such as Google, Yahoo or Microsoft Network, MSN) even when the application provider is not directly connected to the local access network.

The imposition of price discrimination on the provider side of the market and not on the subscriber is a version of two-sided pricing. It is uniquely possible for firms operating within a network structure. Besides traditional networks, such two-sided pricing is also possible for

intermediaries in exchange networks (such as the exchanges themselves). There is presently considerable debate on the legality as well as the efficiency properties of the implementation of such complex pricing strategies by broadband Internet access networks, mainly because of the very considerable market power of such firms.

Residential retail broadband Internet access customers may well have difficulty changing ISPs. Ninety-nine per cent of US households are offered Internet access by at most two firms – a telephone company through DSL and a cable TV company through a cable 'modem' – and many households are facing a monopoly of either cable or DSL. There are also switching costs to residential customers, such as changing equipment. Finally, residential customers are much more affected by contracts that bundle broadband Internet access with other services such as telecommunications and cable television.

As discussed earlier, the Internet under net neutrality separated the network layer from the applications/services layer. This allowed firms to innovate 'at the edge of the network' without seeking approval from network operator(s). The decentralization of the Internet based on net neutrality facilitated innovation resulting in big successes such as Google, MSN, Yahoo, and Skype. Net neutrality also increased competition among the applications and services 'at the edge of the network' which did not need to own a network to compete. Additionally, the existence of network effects on the Internet implies that efficient prices to users on both sides (consumers and applications) should be lower than in a market without network effects. Instead we see an attempt to increase prices that will reduce network effects and innovation.

Abolition of net neutrality raises both horizontal and vertical antitrust issues. To start with horizontal issues, last-mile carriers (who are selling as a duopoly or monopoly to residential consumers) may reduce capacity of 'plain' broadband Internet access service and/or degrade it so that they can establish a 'premium' service for which they intend to charge content/applications providers whose content or application is used by residential subscribers. Coordinated reduction of capacity in 'plain' service is reminiscent of cartel behaviour. In general, the coordinated introduction of price discrimination schemes may reduce output, which would reduce total surplus. Therefore, introduction of coordinated price discrimination may have anti-competitive consequences.

There is also a variety of potentially anti-competitive vertical effects. For example, a carrier may favour its own content or application over the content of a competing carrier or a company that does not have its own network. VOIP provided over broadband Internet competes with traditional circuit-switched service provided by AT&T and Verizon, and could be subject to discrimination. Additionally, both AT&T and Verizon are gearing to distribute video, and could favour their video services over those of others. But the anti-competitive concerns are hardly limited to products and services currently provided by the firms with market power in the access market. The carriers can also leverage market power in broadband access to the content or applications markets through contractual relationships. For example, a carrier can contract with an Internet search engine to put it in 'premium' service while searches using other search engines face considerable delays using 'plain' service. The question that confronts the US Congress in 2007 is whether it should intervene by imposing non-discrimination restrictions or wait instead for antitrust suits to filed and resolved. The crucial role of the Internet in US economic growth argues in favour of pre-emptive restrictions.

Backbone Issues

Backbone networks provide transport and routing services for information packets among high-speed hubs on the Internet. Backbone networks vary in terms of their geographic coverage. There is wide variance of ISPs in terms of their subscriber size and the networks they own. However, irrespective of its size, an ISP needs to interconnect with other ISPs so that its customers can reach all computers/nodes on the Internet. That is, interconnection is necessary to provide the universal connectivity on the Internet which is demanded by users. Internet networks interconnect in two ways: (*a*) private bilateral

interconnection, and (*b*) interconnection at public network access points (NAPs). Private interconnection points and public NAPs are facilities that provide collocation space and a switching platform so that networks are able to interconnect. Interconnection services are complementary to Internet transport. In a sense, the Internet backbone networks are like freeways and the NAPs are like the freeway interchanges.

Internet networks have contracts that govern the terms under which they pay each other for connectivity. Payment takes two distinct forms: (*a*) payment in dollars for 'transit', and (*b*) payment in kind, that is, barter, called 'peering'. Connectivity arrangements among ISPs encompass a seamless continuum, including ISPs that rely exclusively on transit to achieve connectivity, ISPs that use only peering to achieve connectivity, and everything in between. Although there are differences between transit and peering in the specifics of the payments method, and transit includes services to the ISP not provided by peering, these two are essentially alternative payment methods for connectivity. The transport and routing that backbone networks offer do not necessarily differ depending on whether cash (transit) or barter (peering) is used for payment.

Under transit, a network X connects to network Y with a pipeline of a certain size, and pays network Y for allowing X to reach all Internet destinations. Under transit, network X pays Y to reach not only Y and its peers, *but also any other network*, such as network Z by passing through Y, as in the diagram below.

X – – – – – – – – – – – – Y – = – = – = – = – = – Z

Under peering, two interconnecting networks agree not to pay each other for carrying the traffic exchanged between them as long as the traffic originates and terminates in the two networks. In the diagram above, if X and Y have a peering agreement, they exchange traffic without paying each other so long as such traffic terminating on X originates in Y, and traffic terminating on Y originates in X. If Y were to pass to X traffic originating from a network Z that was not a customer of Y, Y would have to pay a transit fee to X (or get paid a transit fee by X), that is, it would not be covered by the peering agreement between X and Y.

Although the networks do not exchange money in a peering arrangement, the price of the traffic exchange is not zero. If two networks X and Y enter into a peering agreement, it means that they agree that the cost of transporting traffic from X to Y and vice versa that is incurred within X is roughly the same as the cost of transporting traffic incurred within Y. These two costs have to be roughly equal if the networks peer, but they are not zero.

It is a commercial decision whether interconnection takes the form of peering or transit payment. Peering is preferred when the cost incurred by X for traffic from X to Y and Y to X is roughly the same as the cost incurred by Y for the same traffic. If not, the networks will use transit. As I explain below, the decision on whether to peer depends crucially on the geographic coverage of the candidate networks.

Generally, peering does not imply that the two networks should have the same size in terms of the numbers of ISPs connected to each network, or in terms of the traffic that the two networks generate. If two networks, X and Y, are similar in terms of the types of users to whom they sell services, the amount of traffic flowing across their interconnection point(s) will be roughly the same, irrespective of the relative size of the networks. For example, suppose that network X has ten ISPs and network has one ISP. If all ISPs have similar features, the traffic flow from X to Y is generally equal to the traffic flow from Y to X.

What determines whether a peering arrangement is efficient for both networks is the *cost* of carrying the mutual traffic within each network. This cost will depend crucially on a number of factors, including the geographic coverage of the two networks. Even if the types of ISPs of the two networks are the same as in the previous example (and therefore the traffic flowing in each direction is the same), the cost of carrying the traffic can be quite different in network X from network Y. For example, network X (with the ten ISPs) may cover a larger geographic area and have significantly higher costs per unit of traffic than network

Y. Then network X would not agree to peer with Y. These differences in costs ultimately would determine the decision to peer (barter) or receive a cash payment for transport.

Where higher costs are incurred by one of two interconnecting networks because of differences in the geographic coverage of the networks, peering would be undesirable from the perspective of the larger network. Similarly, one expects that networks that cover small geographic areas will peer only with each other. Under these assumptions, who peers with whom is a consequence of the extent of a network's geographic coverage, and may not have any particular strategic connotation. In a theoretical model, Milgrom et al. (2000) show how peering can emerge under some circumstances as an equilibrium in a bargaining model between backbones.

Structural conditions for Internet backbone services (ease of expansion and entry) ensure low barriers to entry and expansion, and easy conversion of other transport capacity to Internet backbone capacity. As discussed later, raw transport capacity as well as Internet transport capacity have grown dramatically. Transport capacity is almost a commodity because of its abundance. The business environment for Internet backbone services is competitive. Generally, ISPs buying transport services face flexible transit contracts of relatively short duration. This is reflected in competitive pricing. Economides (2006a) shows that AT&T and MCI had almost identical prices for transit in 1999 when AT&T's backbone business was significantly smaller than MCI's.

ISPs are not locked in by switching costs of any significant magnitude. Thus, ISPs are in good position to change providers in response to any increase in price, and it would be very difficult for a backbone profitably to increase price. Moreover, a large percentage of ISPs has formal agreements that allow them to route packets through several backbone networks and are able to control the way the traffic will be routed (multi-homing).

When an ISP reaches the Internet through multiple backbones, it has additional flexibility in routing its traffic through any particular backbone. A multi-homing ISP can easily reduce or increase the capacity with which it connects to any particular backbone in response to changes in prices of transit. Thus, multi-homing increases the firm-specific elasticity of demand of a backbone provider. Therefore, multi-homing severely limits the ability of any backbone services provider to profitably increase the price of transport. Any backbone increasing the price of transport will face a significant decrease in the capacity bought by multi-homing ISPs.

Large Internet customers also use multiple ISPs, which is called 'customer multi-homing'. They have chosen to avoid any limitation on their ability to switch traffic among suppliers even in the very shortest of runs. Customer multi-homing has similar effects as ISP multi-homing in increasing the firm-specific elasticity of demand of a backbone provider and limiting the ability of any backbone services provider to profitably increase the price of transport.

Like any network, the Internet exhibits network effects. Network effects are present when the value of a good or service to each consumer rises as more consumers use it, everything else being equal – see Economides (1996), Farrell and Saloner (1985), Katz and Shapiro (1985), and Liebowitz and Margolis (1994). In traditional telecommunications networks, an additional customer to the network increases the value of a network connection to all other customers, since each of them can now make an extra call. On the Internet, an additional user potentially

- adds to the information that all others can reach;
- adds to the goods available for sale on the Internet;
- adds one more customer for e-commerce sellers;
- adds to the number of people who can send and receive e-mail or otherwise interact in through the Internet.

Thus, the addition of an extra computer node increases the value of an Internet connection to each connection.

In networks of interconnected networks, there are large social benefits from the interconnection of the networks and the use of common

standards. A number of networks of various ownership structures have harnessed the power of network externalities by using common standards. Examples of interconnected networks of diverse ownership that use common standards include the telecommunications network, the network of fax machines, and the Internet. Despite the different ownership structures in these three networks, the adoption of common standards has allowed each of them to reap huge network-wide benefits.

As the variety and extent of the Internet's offerings expand, and as more customers and more sites join the Internet, the value of a connection to the Internet rises. Because of the high network externalities of the Internet, consumers on the Internet demand universal connectivity, that is, to be able to connect with every website on the Internet and to be able to send electronic mail to anyone. This implies that every network must connect with the rest of the Internet in order to be a part of it. The demand for universal connectivity on the Internet is stronger than the demand of a voice telecommunications customer to reach all customers everywhere in the world. In the case of voice, it may be possible but very unlikely that a customer might buy service from a long-distance company that does not include some remote country because the customer believes that it is very unlikely that he or she would be making calls to that country. On the Internet, however, one does not know where content is located. If company A did not allow its customers to reach region B or customers of a different company C, customers of A would never be able to know or anticipate what content they would be missing. Thus, consumers' desire for Internet universal connectivity is stronger than for voice telecommunications. Additionally, because connectivity on the Internet is two-way, a customer of company A would be losing exposure of his or her content (and the ability to send and receive e-mails) to region B and customers of company C. It would be difficult for customer A to calculate the extent of the losses accrued to him or her from such actions of company A. Thus, again, customers on the Internet require universal connectivity.

In markets with network externalities, firms may create bottleneck power by using proprietary standards. A firm controlling a standard needed by new entrants to interconnect their networks with the network of the incumbent may be in a position to exercise market power (see Economides 2006b). Often a new technology will enter the market with competing incompatible standards. Competition among standards may have the snowball characteristic attributed to network externalities.

Economics literature has established that using network externalities to affect market structure by creating a bottleneck requires three conditions (see Economides 1996; 1989; Farrell and Saloner 1985; Katz and Shapiro 1985):

- networks use proprietary standards;
- no customer needs to reach nodes of or to buy services from more than one proprietary network; and
- customers are captives of the network to which they subscribe and cannot change providers easily and cheaply.

First, without proprietary standards, a firm does not have the opportunity to create the bottleneck. Second, if proprietary standards are possible, the development of proprietary standards by one network isolates its competitors from network benefits, which then accrue to only one network. The value of each proprietary network is diminished when customers need to buy services from more than one network.

Third, the more consumers are captive and cannot easily and economically change providers, the more valuable is the installed base to any proprietary network. I show below that these conditions fail in the context of the Internet backbone.

For example, if universal connectivity were not offered by a backbone network, a customer or its ISP would have to connect with more than one backbone. This would be similar to the period 1895–1930 when a number of telephone companies run disconnected networks. Eventually most of the independent networks were bought by AT&T, which had a dominant long-distance network. The refusal of AT&T to deal and

interconnect with independents was effective for three key reasons: (*a*) AT&T controlled the standards and protocols under which its network ran; (*b*) long-distance service was provided exclusively by AT&T in most of the United States; and (*c*) the cost to a customer of connecting to both AT&T and an independent was high. None of these reasons applies to the Internet. The Internet is based on public protocols. No Internet backbone has exclusive network coverage of a large portion of the United States. Finally, connecting to more than one backbone (multi-homing) is a common practice by many ISPs and does not require big costs. And ISPs can interconnect with each other through secondary peering, as explained below. Thus, the economic factors that allowed AT&T to blackmail independents into submission in the first three decades of the 20th century are reversed in today's Internet backbone, and therefore would not support a profitable refusal to interconnect by any backbone.

The Internet fails to fulfill any of the three necessary conditions stated above under which a network may be able to leverage network externalities and create a bottleneck. First, there are no proprietary standards on the Internet, so the first condition fails. The scenario of standards wars is not at all applicable to Internet transport, where full compatibility, interconnection and inter-operability prevail. For Internet transport, there are no proprietary standards. There is no control of any technical standard by service providers and none is in prospect. Internet transport standards are firmly public property (Kahn and Cerf 1999; Bradner 1999). As a result, any seller can create a network complying with the Internet standards – thereby expanding the network of interconnected networks – and compete in the market.

In fact, the existence and expansion of the Internet and the relative decline of proprietary networks and services, such as CompuServe, can be attributed to the conditions of inter-operability and the tremendous network externalities of the Internet. AOL, CompuServe, Prodigy, MCI and AT&T folded their proprietary electronic mail and other services into the Internet. Microsoft, thought to be the master of exploiting network effects, made the error of developing and marketing the proprietary MSN. After that product failed to sell, Microsoft re-launched the Microsoft Network as an Internet service provider, adhering fully to the public Internet standard. This is telling evidence of the power of the Internet standard and demonstrates the low likelihood that any firm can take control of the Internet backbone by imposing its own proprietary standard.

Second, customers on the Internet demand *universal connectivity*, so the second condition above fails. Users of the Internet do not know in advance what Internet site they may want to contact or to whom they might want to send e-mail. Thus, Internet users demand from their ISPs, and expect to receive, universal connectivity. This is the same expectation that users of telephones, mail and fax machines have: that they can connect to any other user of the network without concern about compatibility, location, or, in the case of telephone or fax, any concern about the manufacturer of the appliance, the type of connection (wireline or wireless) or the owners of the networks over which the connection is made. Because of the users' demand for universal connectivity, ISPs providing services to end users or to websites must make arrangements with other networks so that they can exchange traffic with *any* Internet customer.

Third, there are no 'captive' ISPs on the Internet, so the third condition fails, for a number of reasons:

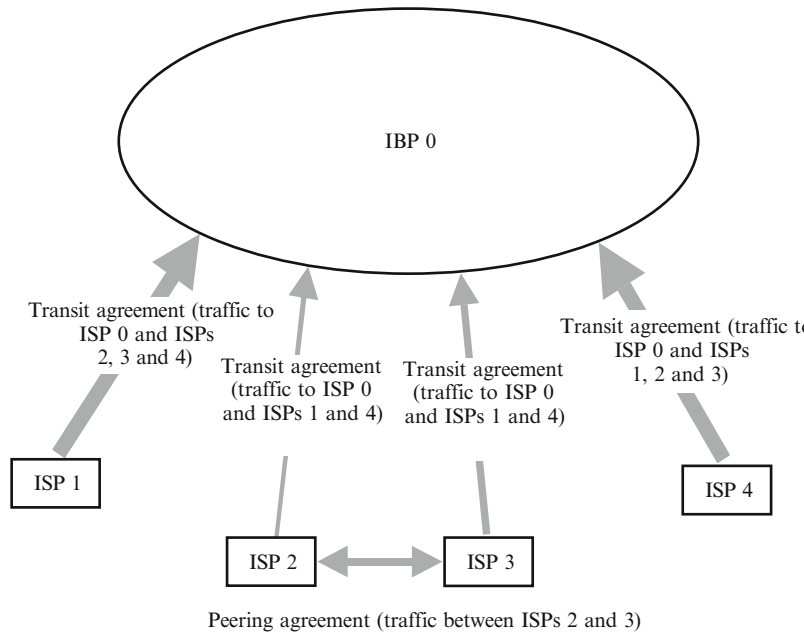
- ISPs can easily and with low cost migrate all or part of their transport traffic to other network providers;
- many ISPs already purchase transport from more than one backbone to guard against network failures and for competitive reasons (ISP 'multi-homing');
- many large websites/providers use more than one ISP for their sites ('customer multi-homing'); and
- competitive pressure from their customers makes ISPs agile and likely to respond quickly to changes in conditions in the backbone market.

Competitive conditions imply that significant price increase, raising rivals' costs or degrading interconnection are unlikely to be profitable on the Internet backbone.

If the large Internet backbone connectivity provider's strategy were to impose equal increases in transport costs on all customers, the response of other backbone providers and ISPs would be to reduce the traffic for which they buy transit from the large Internet backbone provider (IBP) and to instead re-route traffic and purchase more transit from each other. Thus, in response to a price increase by the large Internet backbone connectivity provider, other IBPs and ISPs would reduce the traffic for which they buy transit from the large IBP down to the minimum level necessary to reach ISPs that are *exclusively* connected to the large IBP. All other IBPs and ISPs would exchange all other traffic with each other bypassing the large IBP network.

Figure 2 shows the typical reaction of an increase in the price of a large IBP, and illustrates why the strategy of increasing price is unprofitable. Consider, for example, a situation where, prior to the price increase, four ISPs (1 to 4) purchase transit from IBP 0, which considers increasing its price. Two of these ISPs (ISP 2 and ISP 3) peer with each other. ISP 1 and ISP 4 buy transit capacity for all their traffic to IBP 0 and the other three ISPs. ISP 2 and ISP 3 buy transit capacity for all their traffic to IBP 0, ISP 1 and ISP 4.

Now suppose that IBP 0 increases its transit price. In response, ISP 1 and ISP 4 decide to reduce the traffic for which they buy transit from IBP 0, and instead to reroute some of their traffic and purchase more transit from ISP 2 and ISP 3 respectively. Because of the peering relationship between ISP 2 and ISP 3, all traffic from ISP 1 handed to ISP 2 will reach ISP 3 as well as ISP 4, which is a customer of ISP 3. Similarly, by purchasing transit from ISP 3, ISP 4 can reach all the customers of ISP 1, ISP 2 and ISP 3. Thus, in response to the price increase of IBP each of the ISPs 1, 2, 3 and 4 will reduce the amount of transit purchased from the IBP 0. Specifically, each of the ISPs buys from IBP 0 only capacity sufficient to handle traffic to the customers of network 0. This may lead to a considerable loss in revenues for IBP 0, rendering the price increase unprofitable. The big beneficiaries of the price increase of IBP 0 are peering ISPs 2 and 3, which now start selling transit to ISPs 1 and 4 respectively and become larger networks.



Internet, Economics of the, Fig. 2 Traffic flows between ISPs and a backbone

In response to a price increase by the large IBP, rivals would be able to offer their customers universal connectivity at profitable prices below the large IBP's prices. In the scenario described in the example above, market forces, responding to a price increase by a large network, re-route network traffic so that it is served by rival networks, except for the traffic to and from the ISPs connected exclusively with the large network. The rivals purchase the remaining share from the large IBP in order to provide universal connectivity. Thus, the rivals' blended cost would permit them to profitably offer all transport at prices lower than the large IBP's prices, but above cost.

A direct effect of the increase in price by the large network is that (*a*) ISPs that were originally exclusive customers of the large IBP would shift a substantial portion of their transit business to competitors, and (*b*) ISPs that were not exclusive customers of the large IBP would also shift a significant share of their transit business to competitors' networks, keeping the connection with the large IBP only for traffic for which alternative routes do not exist or for cases of temporary failure of the rivals' networks.

Similarly, degradation of interconnection to all backbones or sequentially one at a time is likely to be unprofitable. Degradation of interconnection to all backbones is clearly dominated by a price increase (since a price increase directly produces additional revenue to the firm, while interconnection degradation does not directly increase revenue), and, as we have shown above, competitive conditions severely limit price increases. Targeted degradation is also unprofitable for a large network that would initiate it for several reasons.

1. ISP clients of the targeted network are likely to switch to third IBP networks that are unaffected by the degradation; it is very unlikely that any will switch to the degrading IBP network because it is itself degraded and cannot offer universal connectivity; there is no demand reward to the large IBP network.
2. Degradation of interconnection hurts all the ISP customers of the targeting IBP network as well, since they lose universal connectivity; these customers of the large network would now be willing to pay less to the large network; this leads to significant revenue and profit loss.
3. After losing universal connectivity, customers of the large IBP network are likely to switch to other networks that are unaffected by degradation and can provide universal connectivity; this leads to even further revenue and profit loss for the degrading network.
4. Multi-homing ISPs would purchase less capacity from the large IBP network, or even terminate their relationship with the large network, which through its own actions sabotages their demand for universal connectivity; this further reduces demand and profits for the degrading network; the same argument applies to multi-homing customers of ISPs.
5. As the large IBP network pursues target after target, its customers face continuous quality degradation while the target's customers face only temporary degradation; this would result in further customer and profit losses for the large IBP network.
6. Prospective victims would seek alternative suppliers in advance of being targeted by the large IBP network; the scheme cannot play out the way it is proposed.
7. The degradation scheme is implausible in its implementation. How large do networks need to be to become serial killers? Why have we not observed this behaviour at all?
8. There is no enduring change to the number of competitors in a market caused by serial degradation in a market with negligible entry barriers; the eliminated rival is likely to be replaced by another.

In conclusion, competition on the Internet backbone is strong, with many carriers and easy entry, and thus presently there are no significant competition concerns for Internet backbone services. However, local broadband access is typically a duopoly or monopoly depending on location. As of 2007, local broadband access networks were proposing to abolish the regime of net neutrality and impose fees on content and applications providers. The legality of this proposed

change is questionable, and imposition of such price discrimination may have adverse consequences for consumers' total surplus.

See Also

► Computer Industry

Bibliography

Bradner, S. 1999. The Internet standards process, revision 3, Network Working Group. Online. Available at ftp://ftp.isi.edu/in-notes/rfc2026.txt. Accessed 24 Jan 2007.

Cremer, J., P. Rey, and J. Tirole. 1998. The degradation of quality and the domination of the internet. Appendix 5 in the *Submission of GTE to the European Union for the merger of MCI with WorldCom.*

Cremer, J., P. Rey, and J. Tirole. 2000. Connectivity in the commercial internet. *Journal of Industrial Economics* 48: 433–472.

Economides, N. 1989. Desirability of compatibility in the absence of network externalities. *American Economic Review* 79: 1165–1181.

Economides, N. 1996. The economics of networks. *International Journal of Industrial Organization* 14: 675–699.

Economides, N. 2006a. The economics of the internet backbone. In *Handbook of telecommunications*, vol. 2, ed. S. Majumdar, I. Vogelsang, and M. Cave. Amsterdam: Elsevier Publishers.

Economides, N. 2006b. Competition policy in network industries: An introduction. In *The new economy and beyond: Past, present and future*, ed. J. Dennis. Cheltenham: Edward Elgar.

European Commission. 1998. *Statement of objections to the MCI WorldCom Merger.*

Farrell, J., and G. Saloner. 1985. Standardization, compatibility, and innovation. *RAND Journal of Economics* 16: 70–83.

Kahn, R.E., and V.G. Cerf. 1999. *What is the internet (and what makes it work)*. Reston: Corporation for National Research Initiatives. Available at http://www.cnri.reston.va.us/what_is_internet.html. Accessed 24 Jan 2007.

Katz, M., and C. Shapiro. 1985. Network externalities, competition and compatibility. *American Economic Review* 75: 424–440.

Liebowitz, S.J., and S.E. Margolis. 1994. Network externality: An uncommon tragedy. *Journal of Economic Perspectives* 8(2): 133–150.

Milgrom, P., B.M. Mitchell, and P. Srinagesh. 2000. Competitive effects of internet peering policies. In *The Internet upheaval – Raising questions, seeking answers in communications policy*, ed. I. Vogelsang and B.M. Compaine. Cambridge, MA/London: MIT Press.

Interpersonal Utility Comparisons

John C. Harsanyi

Abstract

Although we all make interpersonal utility comparisons, many economists and philosophers argue that our limited information about other people's minds renders them meaningless. If they are possible, interpersonal comparisons of utility *differences* must be distinguished from interpersonal comparisons of utility *levels*. Utilitarianism must assume the interpersonal comparability of utility differences to maximize a social welfare function, while Rawls's maximin principle requires interpersonal comparability of utility levels. Adopting an ordinalist or a cardinalist view of utility functions restricts the positions one can consistently take as to interpersonal comparability of utilities.

Keywords

Arrow, K.; Interpersonal utility comparisons; Maximin; Rawls, J.; Robbins, L.; Utilitarianism; Utility: cardinal vs. ordinal; von Neumann–Morgenstern utility function

JEL Classifications

D1

Suppose I am left with a ticket to a Mozart concert I am unable to attend and decide to give it to one of my closest friends. Which friend should I actually give it to? One thing I will surely consider in deciding this is which friend of mine would enjoy the concert *most*. More generally, when we decide as private individuals whom to help, or decide as voters or as public officials who are to receive government help, *one* natural criterion we use is who would derive the greatest benefit, that is, who would derive the *highest utility*, from this help. But to answer this last question we must

make, or at least attempt to make, *interpersonal utility comparisons*.

At the common-sense level, all of us make such interpersonal comparisons. But philosophical reflection might make us uneasy about their meaning and validity. We have direct introspective access only to our *own* mental processes (such as our preferences and our feelings of satisfaction and dissatisfaction) defining our *own* utility function, but have only very indirect information about other people's mental processes. Many economists and philosophers take the view that our limited information about other people's minds renders it impossible for us to make meaningful interpersonal comparisons of utility.

Comparisons of Utility Levels vs. Comparisons of Utility Differences

In any case, if such comparisons are possible at all, then we must distinguish between interpersonal comparisons of utility levels and interpersonal comparisons of utility differences (i.e. utility increments or decrements).

It is one thing to compare the utility level $U_i(A)$ that individual i enjoys (or would enjoy) in situation A, with utility level $U_j(B)$ that another individual j enjoys (or would enjoy) in situation B (where A and B may not refer to the same situation). It is a very different thing to make interpersonal comparisons between utility differences, such as comparing the utility increment

$$\Delta U_i\left(A, A'\right) = U_i\left(A'\right) - U_i(A) \qquad (1)$$

that individual i would enjoy in moving from situation A to situation A', with the utility increment

$$\Delta U_j\left(B, B'\right) = U_j\left(B'\right) - U_j(B) \qquad (2)$$

that individual j would enjoy in moving from B to B'. Either kind of interpersonal comparison might be possible without the other kind being possible (Sen 1970).

Some ethical theories would require one kind of interpersonal comparisons; others would require the other. Thus, *utilitarianism* must assume the interpersonal comparability of utility *differences* because it asks us to maximize a social utility function (social welfare function) defined as the *sum* of all individual utilities. (There are arguments for defining social utility as the *arithmetic mean*, rather than the *sum*, of individual utilities (Harsanyi 1955). But for most purposes – other than analysing population policies – the two definitions are equivalent because if the number of individuals can be taken for a *constant*, then maximizing the sum of utilities is mathematically equivalent to maximizing their arithmetic mean.) Yet, we cannot add different people's utilities unless all of them are expressed in the same utility units; and in order to decide whether this is the case, we must engage in interpersonal comparisons of utility differences. (On the other hand, utilitarianism does not require comparisons of different people's utility levels because it does not matter whether their utilities are measured from comparable zero points or not.)

Likewise, the interpersonal utility comparisons we make in everyday life are most of the time comparisons of utility *differences*. For instance, the comparisons made in our example between the utilities that different people would derive from a concert obviously involve comparing utility differences.

In contrast, the utility-based version of Rawls's *Theory of Justice* (1971) does require interpersonal comparisons of utility *levels*, but does not require comparisons of utility *differences*. This is so because his theory uses the *maximin principle* (he calls it the *difference principle*) in evaluating the economic performance of each society, in the sense of using the well-being of the *worst-off* individual (or the worst-off social group) as its principal criterion. But to decide which individuals (or social groups) are worse off than others he must compare different people's utility levels. (In earlier publications, Rawls seemed to define the worst-off individual as one with the lowest utility level. But in later publications, he defined him as one with the smallest amount of 'primary

goods'. For a critique of Rawls's theory, see Harsanyi 1975).

Ordinalism, Cardinalism and Interpersonal Comparisons

In studying comparisons between the utilities enjoyed by *one* particular individual i, we again have to distinguish between comparisons of utility *levels* and comparisons of utility *differences*. The former would involve comparing the utility levels $U_i(A)$ and $U_i(B)$ that i assigns to two different situations A and B. The latter would involve comparing the utility increment

$$\Delta U_i\left(A, A'\right) = U_i\left(A'\right) - U_i(A) \qquad (3)$$

that i would enjoy in moving from situation A to situation A', with the utility increment

$$\Delta U_i\left(B, B'\right) = U_i\left(B'\right) - U_i(B) \qquad (4)$$

that he would enjoy in moving from B to B'.

If i has a well-defined utility function U_i at all, then he certainly must be able to compare the utility *levels* he assigns to various situations; and such comparisons will have a clear behavioural meaning because they will correspond to the preference and indifference relations expressed by his choice behaviour. In contrast, it is immediately less obvious whether comparing utility differences as defined under (3) and (4) has any economic meaning (but see below).

A utility function U_i permitting meaningful comparisons only between i's utility levels, but not permitting such comparisons between his utility differences, is called ordinal; whereas a utility function permitting meaningful comparisons both between his utility levels and his utility differences is called cardinal.

As is well known, most branches of economic theory use only ordinal utilities. But, as von Neumann and Morgenstern (1947) have shown, cardinal utility functions can play a very useful role in the theory of risk taking. In fact, utility-difference comparisons based on von Neumann–Morgenstern utility functions turn out to have a direct behavioural meaning. For example, suppose that U_i is such a utility function, and let Δ_i^* and Δ_i^{**} be utility differences defined by (3) and by (4). Then, the inequality $\Delta_i^* > \Delta_i^{**}$ will be algebraically equivalent to the inequality

$$\frac{1}{2}U_i\left(A'\right) + \frac{1}{2}U_i(B) > \frac{1}{2}U_i\left(B'\right) + \frac{1}{2}U_i(A). \qquad (5)$$

This inequality in turn will have the behavioural interpretation that i *prefers* an equi-probability mixture of A' and of B to an equi-probability mixture of B' and of A. Of course, once von Neumann–Morgenstern utility functions are used in the theory of risk taking, they become available for possible use also in other branches of economic theory, including welfare economics as well as in ethical investigations. (It has been argued that von Neumann–Morgenstern utility functions have no place in ethics (or in welfare economics) because they merely express people's attitudes toward *gambling*, which has no moral significance (Arrow 1951, p. 10; Rawls 1971, pp. 172 and 323). But see Harsanyi 1984.)

Note that by taking an ordinalist or a cardinalist position, one restricts the positions one can consistently take as to interpersonal comparability of utilities:

(1) An *ordinalist* is logically free to *reject* both types of interpersonal comparisons. Or he may *admit* comparisons of different people's utility *levels*. But he *cannot* admit the interpersonal comparability of utility differences without becoming a cardinalist. (The reason is this. If the utility differences experienced by one individual i are comparable with those experienced by *another* individual j, this will make the utility differences experienced by *one* individual (say) i likewise indirectly comparable with one another, which will enable us to construct a *cardinal* utility function for each individual.)

(2) A *cardinalist* is likewise logically free to *reject* both types of interpersonal

comparisons. Or he may *admit* both. Or else he may admit interpersonal comparisons only for utility *differences*. (Though it is hard to see why anybody might want to reject interpersonal comparisons for utility levels if he admitted them for utility differences.) But he *cannot* consistently admit interpersonal comparisons for utility *levels* while rejecting them for utility *differences*. (This can be verified as follows. If utility levels are interpersonally comparable, then we can find four situations A, A', B, and B' such that $U_i(A) = U_j(B)$ and $U_j(A') = U_j(B')$. But then we can conclude that

$$\Delta_i^* = U_i(A') - U_i(A) = \Delta_j^* = U_i(B') - U_i(B')$$

which means that at least the utility differences Δ_i^* and Δ_j^* are interpersonally comparable. But since U_i and U_j are *cardinal* utility functions, any utility difference Δ_i^{**} experienced by i is comparable with Δ_i^*, and any utility difference Δ_j^{**} experienced by j is comparable with Δ_j^*. Yet this means that *all* utility differences Δ_i^{**} experienced by i are comparable with *all* utility differences Δ_j^{**} experienced by j. Thus, cardinalism together with interpersonal comparability of utility levels *entails* that of utility differences.)

Extended Utility Functions

In what follows, I will use the symbols A_i, B_i, ... to denote the economic and non-economic resources available to individual i in situations A, B, ... Moreover, I will use the symbol A_j to denote an arrangement under which j has the same resources available to him as were available to individual i under arrangement A_i. These entities A_i, B_i, ..., A_j, B_j, ... I will call positions.

Interpersonal utility comparisons would pose no problem if all individuals had the same utility function. For in this case, any individual j could assume that the utility level $U_i(A_i)$ that another individual i would derive from a given position A_i should be the *same* as he himself would derive from a similar position. Thus, j could write simply.

$$U_i(A_i) = U_j(A_j). \tag{6}$$

Of course, in actual fact, the utility of different people are rather *different* because people have different *tastes*, that is, they have different abilities to derive satisfactions from given resource endowments. I will use the symbols R_i, R_j, ... to denote the vectors listing the personal psychological characteristics of each individual i, j, ... that *explain* the differences among their utility functions U_i, U_j, ... Presumably, these vectors summarize the effects that the genetic make-up, the education and the life experience of each individual have on his utility function. This means that any individual j can attempt to assess the utility level $U_i(A_i)$ that another individual j would enjoy in position A_i as

$$U_i(A) = V(A_i, R_i), \tag{7}$$

where the function V represents the psychological laws determining the utility functions U_i, U_j, ... of the various individuals i, j, ... in accordance with their psychological parameters specified by the vectors R_i, R_j, ... Since, by assumption, all differences among the various individuals' utility functions U_i, U_j, ... are fully explained by the vectors R_j, R_j, ..., the function V itself will be the same for all individuals. We will call V an *extended utility function*. (See Arrow 1978; Harsanyi 1977, pp. 51–60; though the basic ideas are contained already in Arrow 1951, pp. 114–15.)

To be sure, we know very little about the psychological laws determining people's utility functions and, therefore, know very little about the true mathematical form of the extended utility function V. This means that, when we try to use Eq. (7), the best we can do is to use our – surely very imperfect – personal *estimate* of V, rather than V itself. As a result, in trying to make interpersonal utility comparisons, we must expect

to make significant errors from time to time – in particular when we are trying to assess the utility functions of people with a very different cultural and social background from our own. But even if our judgements of interpersonal comparisons can easily be mistaken, this does not imply that they are meaningless.

Ordinalists will interpret both the functions U_i and the function V as ordinal utility functions and will interpret (7) merely as a warrant for interpersonal comparisons of utility levels (cf. Arrow 1978). In contrast, cardinalists will interpret all these as *cardinal* utility functions and will interpret (7) as a warrant for *both* kinds of interpersonal comparison (cf. Harsanyi 1977).

Limits to Interpersonal Comparisons

It seems to me that economists and philosophers influenced by *logical positivism* have greatly exaggerated the difficulties we face in making interpersonal utility comparisons with respect to the utilities and the disutilities that people derive from ordinary commodities and, more generally, from the ordinary pleasures and calamities of human life. (A very influential opponent of the possibility of meaningful interpersonal utility comparisons has been Robbins 1932.) But when we face the problem of judging the utilities and the disutilities that other people derive from various *cultural* activities, we do seem to run into very real, and sometimes perhaps even unsurmountable, difficulties. For example, suppose I observe a group of people who claim to derive great aesthetic enjoyment from a very esoteric form of abstract art, which does not have the slightest appeal to me in spite of my best efforts to understand it. Then, there may be no way for me to decide whether the admirers of this art form *really* derive very great and genuine enjoyment from it, or merely *deceive themselves* by claiming that they do.

Maybe in such cases interpersonal comparisons of utility do reach unsurmountable obstacles. But, fortunately, very few of our personal moral decisions and of our public political decisions depend on such exceptionally difficult interpersonal comparisons of utility. (References additional to those listed below will be found in Hammond 1977 and in Suppes and Winet 1955).

See Also

- ▶ Interdependent Preferences
- ▶ Interpersonal Utility Comparisons (New Developments)
- ▶ Pigou, Arthur Cecil (1877–1959)
- ▶ Value Judgements
- ▶ Welfare Economics

Bibliography

Arrow, K.J. 1951. *Social choice and individual values*. 2nd ed. New York: Wiley, 1963.

Arrow, K.J. 1978. Extended sympathy and the possibility of social choice. *Philosophia* 7: 223–237.

Hammond, P.J. 1977. Dual interpersonal comparisons of utility and the welfare economics of income distribution. *Journal of Public Economics* 7: 51–71.

Harsanyi, J.C. 1955. Cardinal utility, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321. Reprinted as ch. 2 of Harsanyi (1977).

Harsanyi, J.C. 1975. Can the maximum principle serve as a basis for morality? A critique of John Rawls' theory. *American Political Science Review* 69: 594–606. Reprinted as ch. 4 of Harsanyi (1977).

Harsanyi, J.C. 1976. *Essays on ethics, social behavior and scientific explanation*. Dordrecht: Reidel.

Harsanyi, J.C. 1977. *Rational behaviour and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.

Harsanyi, J.C. 1984. Von Neumann–Morgenstern utilities, risk taking, and welfare. In *Arrow and the ascent of modern economic theory*, ed. G.R. Feiwel, 545–558. New York: New York University Press.

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

Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.

Sen, A.K. 1970. *Collective choice and social welfare*. San Francisco: Holden-Day.

Suppes, P., and M. Winet. 1955. An axiomatization of utility based on the notion of utility differences. *Management Science* 1: 259–270.

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

Interpersonal Utility Comparisons (New Developments)

Claude d'Aspremont

Abstract

Recent developments on interpersonal utility comparisons rely on various interpretations of 'utility' indicators and combine in various degrees the 'subjective' appreciation of the social states by each individual and their 'objective' evaluation by the ethical observer. In a formal welfarist approach, interpersonal comparisons are specified by invariance conditions on social welfare functionals or on social welfare orderings. Interpersonal comparisons have also been introduced through scoring methods.

Keywords

Bargaining solution (Nash); Bentham, J.; Capabilities; Cardinal utility; Collective preference; Difference principle (J. Rawls); Egalitarian-equivalent allocation; Fair allocation; Happiness; Harsanyi, J. C.; Impossibility theorem; Independence of irrelevant alternatives; Interpersonal utility comparisons; Justice; Leximin; Mill, J. S.; New welfare economics; *Ophélimité*; Ordinal utility; Pareto, V.; Preferences; Primary goods; Rawls, J.; Revealed preference theory; Risk aversion; Sen, A.; Social choice; Social welfare function; Social welfare ordering; Utilitarianism; Utility; Veil of ignorance; Von Neumann and Morgenstern; Welfarism

JEL Classifications

D1

Distributive justice, whether in normative economics or in collective choice theory, can hardly be treated without introducing some interpersonal comparisons. But does this mean considering interpersonal *utility* comparisons? The term 'utility' has received so many different interpretations that the distinguishing mark of the utility approach to the evaluation of *social states* by an ethical observer is simply that it assigns to each member of the collectivity a *unidimensional individual indicator*. These indicators combine in various ways the 'subjective' appreciation of the social states by each individual and their 'objective' evaluation by the observer.

Bentham introduced an interpersonally summable notion of utility based on the objective property of things to procure either pain or pleasure, but it is the subjective reinterpretation given by J.S. Mill, for whom utility means 'pleasure itself' and 'exemption of pain', which has prevailed (Mongin and d'Aspremont 1998). In Harsanyi's (1977) reformulation of utilitarianism, the ethical observer is, with equal chance, any one of the individuals and obeys the rationality conditions of decision-making under risk. The criterion is expected utility computed from the individual von Neumann −Morgenstern (NM) utility functions, but 'corrected' for factual errors and 'censored' for anti-social attitudes.

Pareto has clearly distinguished between the objective notion of *utilité* (in the Bentham sense) and the subjective notion of *ophélimité*, as an ordinal measure of actual preference satisfaction. The latter has become the dominant concept in economics (the only concept in the new welfare economics), forbidding interpersonal utility comparisons and, ultimately, reducing preferences to observable individual choices (revealed preference theory). But Samuelson (1947) insisted that welfare economics cannot avoid ethical and interpersonal assumptions, and Arrow (1951) derived an 'impossibility theorem'. For a set $N = \{1, 2, ..., n\}$ of individuals and a set $X = \{x, y, ...\}$ of (more than two) social states, there is no acceptable *social welfare function*, associating every profile of individual preference orderings with one 'collective' preference ordering of X, and satisfying *weak Pareto* (if all strictly prefer one state to another, so should society) and *independence of irrelevant alternatives* (if individual preferences are modified

except for a subset of several alternatives, then the collective preference should not be modified on this subset). Such a social welfare function can only be *dictatorial*: one individual imposes his strict preference. Since all Arrow's assumptions concern 'preference satisfaction', modifying them and reinterpreting 'utility' involve ethical considerations.

Rawls's (1971) principles of justice are agreed upon in some original negotiation where all irrelevant personal features (including personal conceptions of the good) are ignored. Also chosen behind this 'veil of ignorance' is an 'index of primary goods' defined as an objective indicator of the fundamental resources (except for liberties and access to occupations, preliminarily and equally divided) allocated to each person to promote his own conception of the good. If such an indicator can be called 'utility', it is not in the sense of 'happiness' or 'preference satisfaction'. Desires (however intense) and tastes (however inexpensive) are not relevant per se. A similar view is represented by Sen's (1992) notion of capabilities, that is, the set of doings or 'functionings' available to a person, leading to an 'index of functionings'. What is at stake here, as in other theories concerned by opportunities (for example, Roemer (1996)), are the objectively defined conditions allowing individuals to exercise their freedom.

Extended Sympathy, Social Welfare Functionals and Welfarism

To formally examine the role of interpersonal utility comparisons in social choice, it is usual to start within the framework introduced by Sen (1970), in which the basic ingredient is a utility profile given by a real-valued function U defined on elements (x,i) of the Cartesian product $X \times N$. The function U can be seen as a vector of individual indicators $U_i(x) \equiv U(x, i)$, or as the *extended utility function* of an individual, evaluating from a moral viewpoint what it is to be anyone in any social state (exercising 'extended sympathy' or 'empathy'). Moreover, if for individual i one interprets the name i as designating all the characteristics of i, then one could look at the function U as a *fundamental utility function* (Harsanyi 1977), itself a representation of 'human nature' (Kolm 1972), which would then justify why every individual, when adopting the viewpoint of an ethical observer, should have the same extended utility function (or at least the same fundamental preference). Lack of identity could lead to a dictatorial ethical observer (see Suzumura (1996)).

The fundamental utility approach is also used in econometric estimations to define 'adult-equivalent scales' and different forms of exact aggregation (Blackorby and Donaldson 1991; Christensen et al. 1975; Deaton and Muellbauer 1980). Other measurement techniques, such as that which uses the number of 'just-noticeable-differences' between two alternatives (discussed in (Arrow 1951)) or the 'social indicators' approach using questionnaires about degree of happiness (discussed in (Fleurbaey and Hammond 2004; Hammond 1991)) are differently founded.

The U function is a very flexible informational basis to start with. For every x, we can denote U_x the *utility vector* $(U(x, 1), ..., U(x, n))$ in $\mathbb{R}^N$. Taking all functions U in some domain D determines the set of all admissible utility vectors. Sen's (1970) concept of *social welfare functional* (SWFL) associates every admissible extended utility function U in D with one (collective) preference ordering R_U. We denote I_U and P_U the corresponding indifference and strict preference relations. Using this notation, *Pareto indifference* means '$U_x = U_y$ implies xI_Uy' and *strong Pareto* requires in addition that '$U_x \geq U_y$ and $U_x \neq U_y$ implies $x\ P_Uy$'. Also, Arrow's independence of irrelevant alternatives can be weakened to *binary independence*, whereby for any two functions U and V with equal values on two social states x and y we have $xR_Vy \Leftrightarrow xR_Uy$.

An alternative framework is to define directly a *social welfare ordering* (SWO) denoted R^* on the set of admissible utility vectors. If the set of admissible utility vectors is large enough (for example, equal to $\mathbb{R}^N$), then, under Pareto indifference and binary independence, the two

frameworks coincide: $u = U_x$ and $v = U_y$ implies $uR^*v \Leftrightarrow xR_Uy$. This is called *welfarism* and is an extreme form of consequentialism. All the information required for social evaluation is contained in the final utility values. Under welfarism, strong Pareto (SP) reduces to the condition that $u \geq v$, and $u \neq v$ implies uP^*v (P^* denoting strict collective preference).

Invariance Axioms

Measurement theory (Krantz et al. 1971; Roberts 1980) associates with different measurement scales the associated meaningful statements. We are interested in meaningful statements about intrapersonal and interpersonal comparisons of utility. For instance, the Arrowian informational basis for SWFLs requires that only *intrapersonal level comparisons* are meaningful: $R_U = R_V$ whenever, for every i and all x, y, $U(x, i) \geq U(y, i)$ if and only if $V(x, i) \geq V(y, i)$. Another example (for this and others see Bossert and Weymark (2004)) is to consider meaningful *interpersonal comparisons of utility differences*: $R_U = R_V$ whenever, for all w, x, y, z and all i, j, $U(w, i) - U(x, i) \geq U(y, j) - U(z, j)$ if and only if $V(w, i) - V(x, i) \geq V(y, j) - V(z, j)$.

The more standard way (for example, Sen (1977)) to specify the measurability and comparability properties of 'utility' is to introduce invariance transformations $\phi = (\phi_1, \phi_2, \ldots, \phi_n)$, each ϕ_i being a real-valued function on $\mathbb{R}$. In the Arrowian framework, we get the invariance axiom of *ordinality and non-comparability* (ON): if each ϕ_i is increasing and if for every x, $V(x, i) = \phi_i(U(x, i))$, then $R_U = R_V$. Corresponding to interpersonal comparisons of utility differences, we get *cardinality and unit-comparability* (CU): if each ϕ_i is a positive affine transformation ($\phi_i(u_i) = a_i + bu_i$, with $b > 0$) and if, for all (x,i), $V(x,i) = a_i + bU(x,i)$, then $R_U = R_V$.

But there is a third way of specifying such conditions. We have stated these two axioms as restrictions on SWFLs. Under welfarism, they can be translated into axioms on SWOs, as

$$\begin{aligned}&\text{ON}: \text{for any increasing } \phi_i\text{'s},\\ &\quad uR^*v \Leftrightarrow (\phi_1(u_1), \ldots, \phi_n(u_n))\\ &\times R^*(\phi_1(v_1), \ldots, \phi_n(v_n));\\ &\text{CU}: \text{for any } a_i\text{'s}, b > 0,\\ &\quad uR^*v \Leftrightarrow (a_1 + bu_1, \ldots, a_n + bu_n)\\ &\times R^*(a_1 + bv_1, \ldots, a_n + bv_n).\end{aligned}$$

Invariance axioms determine the informational basis for social evaluation. They should not be considered as purely factual. By specifying the kind of information that a social evaluation can or cannot use, these axioms are taking an ethical stance. But their strong ethical implications are better measured when combined with other axioms. To illustrate, the following axiom allows for (and only for) comparisons of utility differences that are intrapersonal. It is *cardinality and non-comparability* (CN): for any a_i's and positive b_i's, if $V(x, i) = a_i + b_iU(x, i)$, for all (x, i), then $R_U = R_V$. Such a SWFL version of this axiom does not exclude interpersonal utility comparisons. It does allow us to compare *ratios of utility differences* of the sort $(U(w, i) - U(x, i))/(U(y, i) - U(z, i))$ between different individuals, and hence to compare measures of risk aversion in case X is specified as a set of lotteries and each $U_i(x)$ as an NM utility function. But the possibility of such comparisons is erased under welfarism, under which cardinal non-comparability reduces to ordinal non-comparability and, with strong Pareto, implies dictatorship (by Arrow's theorem). Under welfarism, ON becomes equivalent to CN: for any a_i's, positive b_i's,

$$\begin{aligned}&uR^*v \Leftrightarrow (a_1 + b_1u_1, \ldots, a_n + b_nu_n)\\ &\times R^*(a_1 + b_1v_1, \ldots, a_n + b_1v_n).\end{aligned}$$

If CN is replaced by CU and dictatorship is excluded by *anonymity* (any utility vector is socially indifferent to any of its permutations), then the only possibility is the *pure utilitarian* SWO: uR^*v if and only if $\sum_{i \in N} u_i \geq \sum_{i \in N} v_i$ (d'Aspremont and Gevers 1977).

This characterization of utilitarianism is directly related to Harsanyi's aggregation theorem since, under welfarism (and NM preferences), cardinality and unit-comparability become equivalent to NM-independence of the

collective preference ordering (Mongin and d'Aspremont 1998).

By giving priority to liberties and access to occupations, Rawls clearly departs from welfarism, even in a formal sense. However, to allocate other primary goods, a common index $V(x_i)$ is fixed, where x_i is the vector of primary goods to be allocated to individual i. Letting, for an allocation $x = (x_1, \ldots, x_n)$, $U(x, i) = V(x_i)$, we can fall again into the welfarist formal framework. Since the index is common to all, the associated invariance axiom is *ordinality and comparability*

$$OC: \quad \text{for any increasing } \hat{\phi}, \quad uR^*v \Leftrightarrow \left(\hat{\phi}, (u_1), \ldots, \hat{\phi}(u_n)\right) \times R^*\left(\hat{\phi}(v_1), \ldots, \hat{\phi}(v_n)\right).$$

Two other axioms are clearly required by Rawls: *anonymity* and *strong Pareto*, the latter being the reason why equal distribution of all primary goods is not the agreed-upon solution. To any u in $\mathbb{R}^N$, one can associate a (re)ordered vector $u_i(\cdot)$ with same components in $\{v \in \mathbb{R}^N : v_1 \leq v_2 \leq \ldots \leq v_n\}$, the set of ordered utility vectors. *Minimal equity* requires that one should never give priority to the best-off individual over the worst-off. Then, under *separability* (in choosing between two utility vectors the indifferent individuals should not be taken into account), the solution is the 'lexicographic maximin' (*leximin*) SWO : for any u, v in $\mathbb{R}^N$, uP^*v if and only if, for some k, $1 \leq k \leq n$, $u_{i(k)} > v_{i(k)}$, and, $\forall j < k$, $1 \leq j \leq n$, $u_{i(j)} = v_{i(j)}$. Leximin formalizes Rawls's 'difference principle'. Other concepts of opportunity equalizations can be so translated into welfarist terms (see Maniquet (2004)).

From a formal viewpoint, in the preceding result only utility levels are both intrapersonally and interpersonally comparable. If we add the same possibility for utility differences we have *full comparability*, that is

$$FC: \quad \text{for any } a, \ b > 0, \quad uR^*v \Leftrightarrow (a + bu_1, \ldots, a + bu_n) \times R^*(a + bv_1, \ldots, a + bv_n).$$

With this type of invariance and the same other assumptions (Deschamps and Gevers 1978), the SWO R^* can be either leximin or utilitarianism, but in a weak sense (that is $\sum_{i \in N} u_i > \sum_{i \in N} v_i$ implies uP^*v).

Many other invariance axioms can be introduced (for example, (d'Aspremont 1985)). Let us give only two more, *ratio-scale measurability*, *without* or *with interpersonal comparisons* of utility:

$$\text{RN}: \quad \text{for any positive } b_i\text{'s}, \quad \text{and } u, \ v \in \mathbb{R}^N, \quad uR^*v \Leftrightarrow (b_1u_1, \ldots, b_nu_n) \times R^*(b_1v_1, \ldots, b_nv_n);$$

$$\text{RC}: \quad \text{for any positive } b, \quad \text{and } u, \ v \in \mathbb{R}^N, \quad uR^*v \Leftrightarrow b_uR^*bv.$$

With ratio-scale measurability the origin is fixed, so that, under *RN*, utility levels are interpersonally comparable if they are of opposite sign (below or above the zero line). Moreover, under RC, all utility levels and differences are comparable. These axioms are most often applied on a positive domain (denoted $\mathbb{R}^N_{++}$). Under RN, ratios of utilities or percentage changes in utility are interpersonally comparable. If we add SP, then a continuous SWO on $\mathbb{R}^N_{++}$ can only be the Nash bargaining solution with status quo point normalized to zero: for some positive θ_i's, uR^*v if and only if $\prod_{i=1}^n u_i^{\lambda_i} \geq \prod_{i=1}^n v_i^{\lambda_i}$. Under *RC* and *SP* the set of continuous and anonymous SWOs is characterized by all homothetic, increasing, continuous and symmetric functions on $\mathbb{R}^N_{++}$ (for these and other results, see Bossert and Weymark (2004)).

Beyond Welfarism: Scoring and Fair Allocation Rules

Even when 'utility' simply represents actual preference satisfaction, it can be severely adjusted by the ethical observer (or rule designer). This is the case for various voting rules or more generally for *scoring rules*. These rules violate binary independence in one way or another, so that welfarism is excluded. Voting methods, such as Borda's, are

generally not acceptable for social evaluation, but some related rules are better candidates (Moulin 1988). Another 'scoring' method is *relative utilitarianism* (axiomatized by Dhillon and Mertens (1999)). Each $U_i(.)$ is supposed to have both a maximum $U_i^{\max}$ and a minimum $U_i^{\min}$ and to represent a NM preference ordering on X and the observer associates to it a NM utility function $V_i(.)$ ordinally equivalent to $U_i(.)$, and then, through individual affine transformations, defines the scoring function $S_i(x) = \left(V_i(x) - V_i^{\min}\right) / \left(V_i^{\max} - V_i^{\min}\right)$. The score $S_i(x)$ is the same whatever the arbitrarily chosen NM utility function $V_i(.)$ representing the preference ordering underlying the initial utility function $U_i(.)$, and measures of curvatures are preserved. The SWFL F is taken to be pure utilitarianism applied to the scores. It satisfies *ordinality and non-comparability* with respect to the utility functions $U_i(x)$ representing the individual preferences. Since the scores are defined as ratios of utility differences, they could be interpersonally compared, but, because their aggregation is utilitarian, it relies only on interpersonal comparisons of differences of scores.

Other concepts have been proposed in the literature on fair allocations, excluding welfarism in terms of the initial utility functions, but ending up applying some SWFL to some recalibrated utility. An example is the Pareto efficient egalitarian-equivalent allocation concept (Pazner and Schmeidler 1978). To illustrate, let $\overline{\omega}$ in $\mathbb{R}_+^L$ be the vector of total quantities of L goods and X be the set of feasible allocations $(x_1, \ldots x_n)$: each x_i is in $\mathbb{R}_+^L$ and $\sum_{i=1}^n x_i \leq \overline{\omega}$. If $U(x, i) = U_i(x_i)$ is increasing in each argument and continuous then one can define an ordinally equivalent function $V_i(x_i)$ such that $U_i(x_i) = U_i(V_i(x_i)\overline{\omega})$. A Pareto efficient allocation $\hat{x}$ in X is *egalitarian-equivalent* if $U_i(\hat{x}_i) = U_i(V_i(\hat{x}_i)\overline{\omega})$ and $V_i(\hat{x}_i) = V_j(\hat{x}_j)$ for all i,j. As observed in Fleurbaey and Hammond (2004), if individual preferences are convex, $\hat{x}$ can be obtained by applying the leximin SWFL on the $V_i(x_i)$'s. Again, starting with a concept defined in terms of purely ordinal and non-comparable utilities (the U_i's), we end up comparing and equalizing utility levels in terms of the V_i's.

See Also

- Arrow's Theorem
- Bargaining
- Equality of Opportunity
- Fair Allocation
- Interpersonal Utility Comparisons
- Pareto Principle and Competing Principles
- Revealed Preference Theory
- Risk Aversion
- Social Choice
- Social Choice (New Developments)
- Social Welfare Function
- Utility
- Welfare Economics

Bibliography

Arrow, K.J. 1951. *Social choice and individual values*. 2nd ed, 1963. New Haven: Yale University Press.

Blackorby, C., and D. Donaldson. 1991. Adult-equivalence scales, interpersonal comparisons of well-being, and applied welfare economics. In *Interpersonal comparisons of well-being*, ed. J. Elster and J.E. Roemer. Cambridge: Cambridge University Press.

Bossert, W., and J.A. Weymark. 2004. Utility in social choice. In *Handbook of utility theory*, ed. S. Barberà, P. Hammond, and C. Seidl, vol. 2. Dordrecht: Kluwer.

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

d'Aspremont, C. 1985. Axioms for social welfare orderings. In *Social goals and social organizations: Essays in memory of Elisha Pazner*, ed. L. Hurwicz, D. Schmeidler, and H. Sonnenschein. Cambridge: Cambridge University Press.

d'Aspremont, C., and L. Gevers. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44: 199–209.

Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behavior*. Cambridge: Cambridge University Press.

Deschamps, R., and L. Gevers. 1978. Leximin and utilitarian rules: A joint characterization. *Journal of Economic Theory* 17: 143–163.

Dhillon, A., and J.-F. Mertens. 1999. Relative utilitarianism. *Econometrica* 67: 417–498.

Fleurbaey, M., and P.J. Hammond. 2004. Interpersonally comparable utility. In *Handbook of utility theory*, ed. S. Barberà, P.J. Hammond, and C. Seidl, vol. 2. Dordrecht: Kluwer.

Hammond, P.J. 1991. Interpersonal comparisons of utility: Why and how they are and should be made. In *Interpersonal comparisons of well-being*, ed. J. Elster and J.E. Roemer. Cambridge: Cambridge University Press.

I

Harsanyi, J.C. 1977. *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.
Kolm, S.C. 1972. *Justice et Equité*. Paris: CNRS.
Krantz, D., R.D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of measurement*, Additive and polynomial representations. Vol. 1. New York: Academic Press.
Maniquet, F. 2004. On the equivalence between welfarism and equality of opportunity. *Social Choice and Welfare* 23: 1–21.
Mongin, P., and C. d'Aspremont. 1998. Utility theory and ethics. In *Handbook of utility theory*, ed. S. Barberà, P. Hammond, and C. Seidl, vol. 1. Dordrecht: Kluwer.
Moulin, H. 1988. *Axioms of cooperative decision making*. Cambridge: Cambridge University Press.
Pazner, E.A., and D. Schmeidler. 1978. Egalitarian-equivalent allocations: A new concept of economic equity. *Quarterly Journal of Economics* 92: 671–687.
Rawls, J. 1971. *A theory of justice*. Cambridge: Harvard University Press.
Roberts, K.W.S. 1980. Interpersonal comparability and social choice theory. *Review of Economic Studies* 47: 421–439.
Roemer, J.E. 1996. *Theories of distributive justice*. Cambridge: Harvard, University Press.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge: Harvard University Press.
Sen, A.K. 1970. *Collective choice and social welfare*. San Francisco: Holden-Day.
Sen, A.K. 1977. On weights and measures: Informational constraints in social welfare analysis. *Econometrica* 45: 1539–1572.
Sen, A.K. 1992. *Inequality re-examined*. Cambridge: Harvard University Press.
Suzumura, K. 1996. Interpersonal comparisons and the possibility of social choice. In *Social choice re-examined*, ed. K.J. Arrow, A.K. Sen, and K. Suzumura, vol. 2. London: Macmillan.

Intertemporal Choice

Christopher F. Chabris, David I. Laibson and Jonathon P. Schuldt

Abstract

Decisions that have consequences in multiple time periods are intertemporal choices. Individuals typically discount delayed rewards much more than can be explained by mortality effects. The most common discount function is exponential in form, but hyperbolic and quasi-hyperbolic functions seem to explain empirical data better. Individual discount rates may be measured in a variety of ways, subject to important methodological caveats. Higher discount rates are empirically associated with a variety of substance abuse and impulsive conditions, including smoking, alcoholism, cocaine and heroin use, gambling, and risky health behaviours. By contrast, low discount rates may be associated with high cognitive ability.

Keywords

Addiction; Discount factor; Discount rate; Discounted utility; Dynamic consistency; Dynamic inconsistency; Felicity; Generalized hyperbolas; Impatience; Instantaneous utility; Intertemporal choice; Mortality; Naive vs. sophisticated; Neuroeconomics; Normative economics; Positive economics; Preference reversal,; Revealed preference; Salience; Time preference

JEL Classifications

C9

Models of Intertemporal Choice

Most choices require decision-makers to trade-off costs and benefits at different points in time. Decisions with consequences in multiple time periods are referred to as intertemporal choices. Decisions about savings, work effort, education, nutrition, exercise, and health care are all intertemporal choices.

The theory of discounted utility is the most widely used framework for analysing intertemporal choices. This framework has been used to *describe* actual behaviour (positive economics) and it has been used to *prescribe* socially optimal behaviour (normative economics).

Descriptive discounting models capture the property that most economic agents prefer current rewards to delayed rewards of similar magnitude. Such time preferences have been ascribed to a

combination of mortality effects, impatience effects, and salience effects. However, mortality effects alone cannot explain time preferences, since mortality rates for young and middle-aged adults are at least 100 times too small to generate observed discounting patterns.

Normative intertemporal choice models divide into two approaches. The first approach accepts discounting as a valid normative construct, using revealed preference as a guiding principle. The second approach asserts that discounting is a normative mistake (except for a minor adjustment for mortality discounting). The second approach adopts zero discounting (or near-zero discounting) as the normative benchmark.

The most widely used discounting model assumes that total utility can be decomposed into a weighted sum – or weighted integral – of utility flows in each period of time (Ramsey 1928):

$$U_t = \sum_{\tau=0}^{T-t} D(\tau) \cdot u_{t+\tau}.$$

In this representation: U_t is total utility from the perspective of the current period, t; T is the last period of life (which could be infinity for an intergenerational model); $u_{t+\tau}$ is flow utility in period $t + \tau$ ($u_{t+\tau}$ is sometimes referred to as felicity or as instantaneous utility); and $D(\tau)$ is the discount function. If delaying a reward reduces its value, then the discount function weakly *declines* as the delay, τ, *increases*:

$$D'(\tau) \leq 0.$$

Economists normalize $D(0)$ to 1. Economists assume that increasing felicity, $u_{t+\tau}$, weakly increases total utility, U_t. Combining all of these assumptions implies,

$$1 = D(0) \geq D(\tau) \geq D(\tau') \geq 0,$$

where $0 < \tau < \tau'$.

Time preferences are often summarized by the rate at which the discount function declines, $\rho(\tau)$. For differentiable discount functions, the discount rate is defined as

$$\rho(\tau) \equiv -\frac{D'(\tau)}{D(\tau)}.$$

(See Laibson 2003, for the formulae for non-differentiable discount functions.) The higher the discount rate the greater the preference for immediate rewards over delayed rewards.

The discount factor is the inverse of the continuously compounded discount rate.

$\rho(\tau)$. So the discount factor is defined as

$$f(\tau) = \lim_{\Delta \to 0} \left(\frac{1}{1 + \rho(\tau)\Delta} \right)^{1/\Delta} = e^{-\rho}(\tau).$$

The lower the discount factor the greater the preference for immediate rewards over delayed rewards.

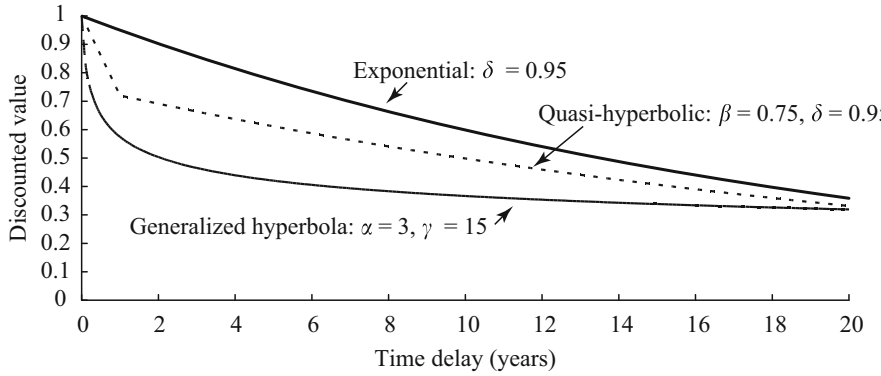
The most commonly used discount function is the exponential discount function:

$$D(\tau) = \delta^{\tau},$$

with $0 < \delta < 1$. For the exponential discount function, the discount rate is independent of the horizon, τ. Specifically, the discount rate is $-\ln(\delta)$ and the discount factor is δ. Figure 1.

The exponential discount function also has the property of dynamic consistency: preferences held at one point in time do not change with the passage of time (unless new information arrives). For example, consider the following investment opportunity: pay a utility cost of C at date $t = 2$ to reap a utility benefit of B at date $t = 3$. Suppose that this project is viewed from date $t = 1$ and judged to be worth pursuing. Hence, $-\delta C + \delta^2 B > 0$. Imagine that a period of time passes, and the agent reconsiders the project from the perspective of date $t = 2$. Now the project is still worth pursuing, since $-C + \delta B > 0$. To prove that this is true, note that the new expression is equal to the old expression multiplied by $1/\delta$. Hence, the $t = 1$ preference to complete the project is preserved at date $t = 2$. The exponential discount function is the *only* discount function that generates dynamically consistent preferences.

Despite its many appealing properties, the exponential discount function fails to match



Intertemporal Choice, Fig. 1 Three calibrated discount functions

several empirical regularities. Most importantly, a large body of research has found that measured discount functions decline at a higher rate in the short run than in the long run. In other words, people appear to be more impatient when they make short-run trade-offs – today vs. tomorrow – than when they make long-run trade-offs – day 100 vs. day 101. This property has led psychologists (Herrnstein 1961; Ainslie 1992; Loewenstein and Prelec 1992) to adopt discount functions in the family of generalized hyperbolas:

$$D(\tau) = (1 + \alpha\tau)^{-\gamma/\alpha}.$$

Such discount functions have the property that the discount rate is higher in the short run than in the long run. Particular attention has been paid to the case in which $\gamma = \alpha$, implying that $D(\tau) = (1 + \alpha\tau)^{-1}$.

Starting with Strotz (1956), economists have also studied alternatives to exponential discount functions. The majority of economic research has studied the quasi-hyperbolic discount function, which is usually defined in discrete time:

$$D(\tau) = \left\{ \begin{array}{ll} 1 & \text{if } \tau = 0 \\ \beta \cdot \delta^{\tau} & \text{if } \tau = 1, 2, 3, \ldots \end{array} \right\}.$$

This discount function was first used by Phelps and Pollak (1968) to study intergenerational discounting. Laibson (1997) subsequently applied this discount function to intra-personal decision problems. When $0 < \beta < 1$ and $0 < \delta < 1$ the quasi-hyperbolic discount function has a high short-run discount rate and a relatively low long-run discount rate. The quasi-hyperbolic discount function nests the exponential discount function as a special case ($\beta = 1$). Quasi-hyperbolic time preferences are also referred to as 'present-biased' and 'quasi-geometric'.

Like other non-exponential discount functions, the quasi-hyperbolic discount function implies that intertemporal preferences are not dynamically consistent. In other words, the passage of time may change an agent's preferences, implying that preferences are dynamically inconsistent. To illustrate this phenomenon, consider an investment project with a cost of 6 at date $t = 2$ and a delayed benefit of 8 at date $t = 3$. If $\beta = 1/2$ and $\delta = 1$ (see Akerlof 1991), this investment is desirable from the perspective of date $t = 1$. The discounted value is positive:

$$\beta(-6 + 8) = \frac{1}{2}(-6 + 8) = 1.$$

However, the project is undesirable from the perspective of date 2. Judging the project from the $t = 2$ perspective, the discounted value is negative:

$$-6 + \beta(8) = -6 + \frac{1}{2}(8) = -2.$$

This is an example of a preference reversal. At date $t = 1$ the agent prefers to do the project at $t = 2$.

At date $t = 2$ the agent prefers not to do the project. If economic agents foresee such preference reversals they are said to be sophisticated and if they do not foresee such preference reversals they are said to be naive (Strotz 1956). O'Donoghue and Rabin (2001) propose a generalized formulation in which agents are partially naive: the agents have an imperfect ability to anticipate their preference reversals.

Many different microfoundations have been proposed to explain the preference patterns captured by the hyperbolic and quasi-hyperbolic discount functions. The most prominent examples include temptation models and dual-brain neuroeconomic models (Bernheim and Rangel 2004; Gul and Pesendorfer 2001; McClure et al. 2004; Thaler and Shefrin 1981). However, both the properties and mechanisms of time preferences remain in dispute.

Individual Differences in Measured Discount Rates

Numerous methods have been used to measure discount functions. The most common technique poses a series of questions, each of which asks the subject to choose between a sooner, smaller reward and a later, larger reward. Usually the sooner, smaller reward is an *immediate* reward. The sooner and later rewards are denominated in the same goods, typically amounts of money or other items of value. For example: 'Would you rather have $69 today, or $85 in 91 days?' The subject's discount rate is inferred by fitting one or more of the discount functions described in the previous section to the subject choices. Most studies assume that the utility function is linear in consumption. Most studies also assume no intertemporal fungibility – the reward is assumed to be consumed the moment it is received. Many factors may confound the analysis in such studies, leading numerous researchers to express scepticism about the conclusions generated by laboratory studies. Table 1 provides a summary of such critiques.

Discount functions may also be inferred from field behaviour, such as consumption, savings, asset allocation, and voluntary adoption of forced-savings technologies (Angeletos et al. 2001; Shapiro 2005; Ashraf et al. 2006). However, field studies are also vulnerable to methodological critiques. There is currently no methodological gold standard for measuring discount functions.

Existing attempts to measure discount functions have reached seemingly conflicting conclusions (Frederick, Lowenstein and O'Donoghue, Frederick et al. 2003). However, the fact that different methods and samples yield different estimates does not rule out consistent individual differences. Dozens of empirical studies have explored the relationship between individuals' estimated discount rates and a variety of behaviours and traits. A significant subset of this literature has focused on delay discounting and behaviour in clinical populations, most notably drug users, gamblers, and those with other impulsivity-linked psychiatric disorders (see Reynolds 2006, for a review). Other work has explored the relationship between discounting and traits such as age and cognitive ability. Table 2 summarizes representative studies.

Smoking A number of investigations have explored the relationship between cigarette smoking and discounting, together providing strong evidence that cigarette smoking is associated with higher discount rates (Baker et al. 2003; Bickel et al. 1999; Kirby and Petry 2004; Mitchell 1999; Ohmura et al. 2005; Reynolds et al. 2004).

Excessive Alcohol Consumption While the association with alcoholism has received relatively little attention, the available data suggest that problematic drinking is associated with higher discount rates. Heavy drinkers have higher discount rates than controls (Vuchinich and Simpson 1998), active alcoholics discount rewards more than abstinent alcoholics, who in turn discount at higher rates than controls (Petry 2001a), and detoxified alcohol-dependents have higher discount rates than controls (Bjork et al. 2004).

Illicit Drug Use Recent studies document a positive association between discount rates and drug

Intertemporal Choice, Table 1 Potential confounds that may arise in attempts to measure discount rates in laboratory studies

Factor	Description
Unreliability of future rewards	A subject may prefer an earlier reward because the subject thinks she is unlikely to actually receive the later reward. For example, the subject may perceive an experimenter as unreliable.
Transaction costs	A subject may prefer an immediate reward because it is paid in cash, whereas the delayed reward is paid in a form that generates additional transaction costs. For example, a delayed reward may need to be collected, or it may arrive in the form of a cheque that needs to be cashed.
Hypothetical rewards	A subject may not reveal her true preferences if she is asked hypothetical questions instead of being asked to make choices with real consequences. However, researchers who have directly compared real and hypothetical rewards have concluded that this difference does not arise in practice (Johnson and Bickel 2002).
Investment versus consumption	Some subjects may interpret a choice in a discounting experiment as an *investment* decision and not a decision about the timing of consumption. For example, a subject might reason that a later, larger reward is superior to a sooner, smaller reward as long as the return for waiting is higher than the return available in financial markets.
Consumption versus receipt	Rewards, especially large ones, may not be consumed at the time they are received. For example, a $500 reward is likely to produce a stream of higher consumption, not a lump of consumption at the date of receipt. Such effects may explain why large-stake experiments are associated with less measured discounting than small-stake experiments
Curvature of utility function	A subject may prefer a sooner, smaller reward to a later, larger reward if the subject expects to receive other sources of income at that later date. In general, a reward may be worth less if it is received during a period of relative prosperity.
Framing effects	The menu of choices or the set of questions may influence the subject's choices. For example, if choices between $1.00 now and delayed amounts ranging between $1.01 and $1.50 were offered, subjects may switch preference from early to later rewards at an interior threshold – for example $1.30. However, if choices between $1.00 and delayed amounts ranging between $1.51 and $2.00 were offered, the switch might happen at a much higher threshold – for example $1.70 – implying a much higher discount rate.
Demand characteristics	Procedures for estimating discount rates may bias subject responses by implicitly guiding their choices. For example, the phrasing of an experimental question can imply that a particular choice is the right or desired answer (from the perspective of the experimenter).

use for a variety of illicit drugs, most notably cocaine, crack-cocaine, heroin and amphetamines (Petry 2003; Coffey et al. 2003; Bretteville-Jensen 1999; Kirby and Petry 2004).

Gambling Pathological gamblers have higher discount rates than controls, both in the laboratory (Petry 2001b) and in a more natural setting (Dixon et al. 2003), and among a population of gambling and non-gambling substance abusers (Petry and Casarella 1999). Moreover, Alessi and Petry (2003) report a significant, positive relationship between a gambling severity measure and the discount rate within a sample of problem gamblers. Petry (2001b) finds that gambling frequency during the previous 3 months correlates positively with discount rate.

Age Patience appears to increase across the lifespan, with the young showing markedly less patience than middle-aged and older adults (Green et al. 1994; Green et al. 1996; Green et al. 1999). Read and Read (2004) report that older adults (mean age = 75) are the most patient age group when delay horizons are only 1 year. However, this study also finds that older adults are the *least* patient group when delay horizons are from three to ten years. This reversal probably reflects the fact that 75-year-olds face significant mortality/disability risk at horizons of three to ten years.

Cognitive Ability Kirby et al. (2005) report that discount rates are correlated negatively with grade point average in two college samples. Benjamin et al. (2006) find an inverse relationship between

Intertemporal Choice, Table 2 Representative empirical studies linking estimated discount rates for monetary rewards to various individual behaviours and traits

Variable	Study	*N*	Discount rate findings
Nicotine	Bickel et al. (1999)*	66	Current smokers > never-smokers and ex-smokers
Alcohol	Bjork et al. (2004)	160	Abstinent alcohol-dependent subjects > controls
Cocaine	Coffey et al. (2003)*	25	Crack-dependent subjects > matched controls[a]
Heroin	Kirby et al. (1999)	116	Heroin addicts > age-matched controls
Gambling	Petry (2001b)*	86	Pathological gamblers[b] > controls
Risky Behaviour	Odum et al. (2000)*	32	Heroin addicts agreeing to share needle in a hypothetical scenario > non-agreeing
Age	Green et al. (1994)*	36	Children > young adults > older adults
Psychiatric disorders	Crean et al. (2000)	24	'High risk' patients [c] > 'low risk' patients
Cognitive ability	Benjamin et al. (2006)	92	Low scorers on standardized mathematics test > high scorers

Notes: N = total number of participants in study
*These studies used hypothetical rewards; others used real rewards
[a]Results based on those choices falling within the delay range of 1 week to 25 years. Overall analyses including shorter delays (5 min to 5 days) also revealed the same effect, but with smaller magnitude
[b]Gamblers with comorbid substance abuse disorders showed a greater effect than gamblers without such disorders
[c]'High risk' patients were those diagnosed with disorders carrying high risk for impulsive behaviour, according to *DSM-IV* criteria, such as patients with borderline personality disorder, bipolar disorder, and substance abuse disorders

individual discount rates and standardized (mathematics) test scores for Chilean high school students. Silva and Gross (2004) show that students scoring in the top third of their introductory psychology course have lower discount rates than those scoring in the middle and lower thirds. Frederick (2005) shows that participants scoring high on a 'cognitive reflection' problem-solving task demonstrate more patient intertemporal choices (for a variety of rewards) than those scoring low. Finally, in a sample of smokers, Jaroni et al. (2004) report that participants who did not attend college had higher discount rates than those attending at least some college.

All of these empirical regularities are consistent with the neuroeconomic hypothesis that prefrontal cortex is essential for patient (forward-looking) decision-making (McClure et al. 2004). This area of the brain is slow to mature, is critical for general cognitive ability (Chabris 2007), and is often found to be dysfunctional in addictive and other psychiatric disorders.

More research is required to clarify the cognitive and neurobiological bases of intertemporal preferences. Future research should evaluate the usefulness of measured discount functions in predicting real-world economic decisions (Ashraf et al. 2006). Finally, ongoing research should improve the available methods for measuring intertemporal preferences.

See Also

▶ Time Preference

Bibliography

Ainslie, G. 1992. *Picoeconomics*. New York: Cambridge University Press.
Akerlof, G.A. 1991. Procrastination and obedience. *American Economic Review* 81: 1–19.
Alessi, S.M., and N.M. Petry. 2003. Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behavioural Processes* 64: 345–354.
Angeletos, G.-M., D.I. Laibson, A. Repetto, J. Tobacman, and S. Weinberg. 2001. The hyperbolic consumption model: Calibration, simulation, and empirical evaluation. *Journal of Economic Perspectives* 15(3): 47–68.
Ashraf, N., D.S. Karlan, and W. Yin. 2006. Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *Quarterly Journal of Economics* 121: 673–697.
Baker, F., M.W. Johnson, and W.K. Bickel. 2003. Delay discounting in current and never-before cigarette smokers: Similarities and differences across

commodity, sign, and magnitude. *Journal of Abnormal Psychology* 112: 382–392.

Benjamin, D.J., Brown, S.A. and Shapiro, J.M. 2006. Who is 'behavioral'? Cognitive ability and anomalous preferences. Unpublished manuscript, University of Michigan.

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

Bickel, W.K., A.L. Odum, and G.J. Madden. 1999. Impulsivity and cigarette smoking: Delay discounting in current, never, and ex-smokers. *Psychopharmacology* 146: 447–454.

Bjork, J.M., D.W. Hommer, S.J. Grant, and C. Danube. 2004. Impulsivity in abstinent alcohol-dependent patients: Relation to control subjects and type 1−/type 2-like traits. *Alcohol* 34: 133–150.

Bornovalova, M.A., S.B. Daughters, G.D. Hernandez, J.B. Richards, et al. 2005. Differences in impulsivity and risk-taking propensity between primary users of crack cocaine and primary users of heroin in a residential substance-use program. *Experimental and Clinical Psychopharmacology* 13: 311–318.

Bretteville-Jensen, A.L. 1999. Addiction and discounting. *Journal of Health Economics* 18: 393–407.

Chabris, C.F. 2007. Cognitive and neurobiological mechanisms of the law of general intelligence. In *Integrating the mind*, ed. M.J. Roberts. Hove: Psychology Press.

Coffey, S.F., G.D. Gudleski, M.E. Saladin, and K.T. Brady. 2003. Impulsivity and rapid discounting of delayed hypothetical rewards in cocaine-dependent individuals. *Experimental and Clinical Psychopharmacology* 11: 18–25.

Crean, J.P., H. de Wit, and J.B. Richards. 2000. Reward discounting as a measure of impulsive behavior in a psychiatric outpatient population. *Experimental and Clinical Psychopharmacology* 8: 155–162.

Dixon, M.R., J. Marley, and E.A. Jacobs. 2003. Delay discounting by pathological gamblers. *Journal of Applied Behavior Analysis* 36(4): 449–458.

Field, M., M. Santarcangelo, H. Sumnall, A. Goudie, et al. 2006. Delay discounting and the behavioural economics of cigarette purchases in smokers: The effects of nicotine deprivation. *Psychopharmacology* 186: 255–263.

Frederick, S. 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives* 19(4): 24–42.

Frederick, S., G. Loewenstein, and T. O'Donoghue. 2003. Time discounting and time preference: A critical review. In *Time and decision: Economic and psychological perspectives on intertemporal choice*, ed. G. Loewenstein, D. Read, and R. Baumeister. New York: Sage.

Giordano, L.A., W.K. Bickel, G. Loewenstein, E.A. Jacobs, et al. 2002. Mild opioid deprivation increases the degree that opioid-dependent outpatients discount delayed heroin and money. *Psychopharmacology* 163: 174–182.

Green, L., A.F. Fry, and J. Myerson. 1994. Discounting of delayed rewards: A life-span comparison. *Psychological Science* 5: 33–37.

Green, L., J. Myerson, and P. Ostazewski. 1999. Discounting of delayed rewards across the life span: Age differences in individual discounting functions. *Behavioural Processes* 46: 89–96.

Green, L., J. Myerson, D. Lichtman, S. Rosen, and A. Fry. 1996. Temporal discounting in choice between delayed rewards: The role of age and income. *Psychology and Ageing* 11: 79–84.

Gul, F., and W. Pesendorfer. 2001. Temptation and self-control. *Econometrica* 69: 1403–1435.

Herrnstein, R.J. 1961. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior* 4: 267–272.

Hinson, J.M., T.L. Jameson, and P. Whitney. 2003. Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 298–306.

Holt, D.D., L. Green, and J. Myerson. 2003. Is discounting impulsive? Evidence from temporal and probability discounting in gambling and non-gambling college students. *Behavioural Processes* 64: 355–367.

Jaroni, J.L., S.M. Wright, C. Lerman, and L.H. Epstein. 2004. Relationship between education and delay discounting in smokers. *Addictive Behaviors* 29: 1171–1175.

Johnson, M.W., and W.K. Bickel. 2002. Within-subject comparison of real and hypothetical money rewards in delay discounting. *Journal of the Experimental Analysis of Behavior* 77: 129–146.

Kirby, K.N., and N.M. Petry. 2004. Heroin and cocaine abusers have higher discount rates for delayed rewards than alcoholics or non-drug-using controls. *Addiction* 99: 461–471.

Kirby, K.N., N.M. Petry, and W.K. Bickel. 1999. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *Journal of Experimental Psychology: General* 128: 78–87.

Kirby, K.N., G.C. Winston, and M. Santiesteban. 2005. Impatience and grades: Delay-discount rates correlate negatively with college GPA. *Learning and Individual Differences* 15: 213–222.

Laibson, D. 2003. *Intertemporal decision making. In encyclopedia of cognitive science*. London: Nature Publishing Group.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112: 443–477.

Loewenstein, G., and D. Prelec. 1992. Anomalies in intertemporal choice: Evidence and an interpretation. *Quarterly Journal of Economics* 107: 573–597.

Madden, G.J., A.M. Begotka, B.R. Raiff, and L.L. Kastern. 2003. Delay discounting of real and hypothetical rewards. *Experimental and Clinical Psychopharmacology* 11: 139–145.

McClure, S.M., D.I. Laibson, G. Loewenstein, and J.D. Cohen. 2004. Separate neural systems value immediate and delayed monetary rewards. *Science* 306: 503–507.

Mitchell, S.H. 1999. Measures of impulsivity in cigarette smokers and non-smokers. *Psychopharmacology* 146: 455–464.
Mitchell, S.H. 2004. Effects of short-term nicotine deprivation on decision-making: Delay, uncertainty, and effort discounting. *Nicotine & Tobacco Research* 6: 819–828.
O'Donoghue, T., and M. Rabin. 2001. Choice and procrastination. *Quarterly Journal of Economics* 116: 121–160.
Odum, A.L., G.J. Madden, G.J. Badger, and W.K. Bickel. 2000. Needle sharing in opioid-dependent outpatients: Psychological processes underlying risk. *Drug and Alcohol Dependence* 60: 259–266.
Ohmura, Y., T. Takahashi, and N. Kitamura. 2005. Discounting delayed and probabilistic monetary gains and losses by smokers of cigarettes. *Psychopharmacology* 182: 508–515.
Ortner, C.N.M., T.K. MacDonald, and M.C. Olmstead. 2003. Alcohol intoxication reduces impulsivity in the delay-discounting paradigm. *Alcohol and Alcoholism* 38: 151–156.
Petry, N.M. 2001a. Delay discounting of money and alcohol in actively using alcoholics, currently abstinent alcoholics, and controls. *Psychopharmacology* 154: 243–250.
Petry, N.M. 2001b. Pathological gamblers, with and without substance use disorders, discount delayed rewards at high rates. *Journal of Abnormal Psychology* 3: 482–487.
Petry, N.M. 2003. Discounting of money, health, and freedom in substance abusers and controls. *Drug and Alcohol Dependence* 71: 133–141.
Petry, N.M., and T. Casarella. 1999. Excessive discounting of delayed rewards in substance abusers with gambling problems. *Drug and Alcohol Dependence* 56: 25–32.
Phelps, E.S., and R.A. Pollak. 1968. On second-best national saving and game-equilibrium growth. *Review of Economic Studies* 35: 185–199.
Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–549.
Read, D., and N.L. Read. 2004. Time discounting over the lifespan. *Organizational Behavior and Human Decision Processes* 94: 22–32.
Reynolds, B. 2006. A review of delay-discounting research with humans: Relations to drug use and gambling. *Behavioural Pharmacology* 17: 651–667.
Reynolds, B., J.B. Richards, K. Horn, and K. Karraker. 2004. Delay discounting and probability discounting as related to cigarette smoking status in adults. *Behavioral Processes* 65: 35–42.
Shapiro, J.M. 2005. Is there a daily discount rate? Evidence from the food stamp nutrition cycle. *Journal of Public Economics* 89: 303–325.
Silva, F.J., and T.F. Gross. 2004. The rich get richer: Students' discounting of hypothetical delayed rewards and real effortful extra credit. *Psychonomic Bulletin & Review* 11: 1124–1128.
Strotz, R.H. 1956. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23: 165–180.
Thaler, R.H., and H.M. Shefrin. 1981. An economic theory of self-control. *Journal of Political Economy* 89: 392.
Vuchinich, R.E., and C.A. Simpson. 1998. Hyperbolic temporal discounting in social drinkers and problem drinkers. *Experimental and Clinical Psychopharmacology* 6: 292–305.

Intertemporal Equilibrium and Efficiency

E. Malinvaud

Abstract

A clear formalization of intertemporal equilibrium not only aids the fundamental conceptualization of economic activity but should also lead to comparative statics properties, which, dealing with intertemporal equilibria, have also been called 'comparative dynamics properties'. Particular importance has been given to the question of knowing how the interest rate changes from one stationary equilibrium to another when some specific change is being brought to its exogenous determinants. The theory of the optimum allocation of resources can likewise be transposed to the intertemporal framework. Applications of these properties may give insights on the evolution of prices through time.

Keywords

Austrian economics; Capital accumulation; Capital theory; Comparative statics; Competitive temporary equilibria; Decision criteria; Exhaustible resources; Golden rule; Impatience; Implicit contracts; Information aggregation; Information exchange; Intertemporal competitive equilibrium; Intertemporal decisions; Intertemporal efficiency; Intertemporal equilibrium; Irreversibilities; Joint production; Keynesian consumption function; Labour market contracts; Overcapitalization; Overlapping generations models; Pareto efficiency; Period of production; Production possibility set; Productivity of investment; Proportional growth;

Rate of interest; Rationing; Reswitching of technique; Samuelson–Leontief technology; Shadow discount rate; Stationary equilibrium; *Temporary equilibrium*; Trade cycle; Unbounded horizons; Uncertainty

JEL Classifications
D9

People, corporations and governments take decisions for the future. What kind of consistency exists between these decisions? What role does the price system play in this respect? Is the resulting evolution efficient? How can economic organization be improved in order to permit a more satisfactory growth?

Confronted with such huge questions, economists have often answered quickly. Even when attention is limited to formal theory, which this article exclusively considers, many statements can be found which, taken as valid for a time, were later disproved. They had been obtained on special models and too easily given a broad validity. Indeed, the preliminary step should have been to find a general formal representation of economic activity through time, but this step was not given sufficient attention until the late 19th century (Böhm-Bawerk 1888; Fisher 1907). The central model with reference to which the whole theory can be built and developed clearly emerged only in the 1950s.

A survey on the subject must then start from first principles and note which major features of reality are still today neglected in main-stream theory. The significance of the most far-reaching results and the importance of some big question marks will then have to be assessed.

Intertemporal Decisions

Households save for future consumption, employees work overtime so as to have enough to enjoy their vacation, students strive to get a diploma so as to hold good jobs later, parents want to leave bequest to their children. Firms produce to inventories in the expectation of future sales, recruit and train staff that will later improve their competitiveness, install equipment to be used for many years, build new factories.

The main theories dealing with intertemporal economic problems see such decisions as parts of plans that the relevant agents make for all their future activities. Any household, for instance, is assumed not only to decide its present supply of labour and demand for goods, but also simultaneously to choose its plan for the labour to be later supplied and the goods to be later consumed, and this up to the end of its existence.

The notion of this plan can in principle be made richer by taking *uncertainties* into account; the future decisions are then conditional on events to be later observed, but they are already specified for all conceivable combinations of events. In principle again the structure of the plan must then depend on the structure of the *information* that the agent will receive. In the main intertemporal theories these complications coming from uncertainties and information are, however, neglected, so that the concept of a plan does not appear to be unduly abstract. When the relevance of these theories is assessed, one has to wonder about the consequences of the simplification, as will be seen in the sequel.

Analysis of intertemporal behaviour can adopt the familiar approach: the constraints to which the plan is subject and the objectives that it strives to achieve must be identified; then the optimization problem is solved. The purest of all theories simply transpose the classical analysis of consumer and producer behaviour (Debreu 1959). They assume the existence of a full system of discounted prices, with one such price for each commodity at each present or future date, a price at which agents will be able to buy or sell as much of this commodity as they may wish. They then directly reinterpret as follows the constraints and objectives that static atemporal theories made familiar.

As between the many plans that he can think of, a consumer is assumed to have a system of preferences that is often conveniently represented by a utility function, whose argument is a consumption vector with as many components as there are commodities and dates. A budget

constraint requires that the discounted value of the consumption vector does not exceed a given amount, the consumer initial wealth.

The chosen plan maximizes the utility function subject to the budget constraint. It then follows that the consumption of the various commodities (and the supply of labour) depend on what are the discounted prices and the initial wealth. The present saving of the consumer may be said to be equal to the interest income earned on his initial wealth *minus* the value of his present consumption (labour income appears negatively in this value). It is immaterial in this theory to know how saving is invested. Hence, the consumption plan and the resulting saving plan are seen as involving the whole future *life cycle* of the consumer (Modigliani and Brumberg 1954).

The plan of a producer is subject only to the constraints that technology imposes. The producer acts as a price taker. His objective is to maximize the discounted value of the plan. It follows that demand for inputs and supply of outputs are functions of the discounted prices. The balance between the value of present outputs and present inputs gives the financial surplus if positive or requirement if negative; this is subject to no direct constraint.

Such a theory of consumer and producer behaviour does not claim to apply to all problems concerning this behaviour. Clearly, analysis of the firm in particular must usually go far beyond the stylized description given above, even simply when investment behaviour is being studied (Nickell 1978). But the theory is supposed to be appropriate for fitting into the discussion of the broad questions raised by intertemporal equilibrium and efficiency.

Even when it is so circumscribed, the intent cannot be considered as fully achieved. Significant limitations must be kept in mind, since they may forbid application of the theory to some of the problems raised by equilibrium and efficiency over time; indeed, some of these limitations have been the motivation for theoretical developments that will not be discussed at length here, but must be mentioned.

Full knowledge of the system of discounted prices for purchases of sales at all relevant future dates is of course an abstraction. Forward prices exist for only a few basic commodities and a limited horizon. Whereas the interest rates at which one can borrow or lend for more or less long durations are fairly well defined, with non-negligible transaction costs and fiscal interference, however, prices that will apply to future transactions have to be forecast by the agents. The uncertainties that their forecast necessarily contains are neglected. Among the many consequences of this major simplification, one particularly notes that it rules out fundamental problems concerning the characterization of decision criteria of business firms (Drèze 1982).

Constraints on individual choices are also reduced to a minimum. No consideration is given to quantitative constraints, such as those following from mass unemployment on individuals looking for jobs or from business depression on firms looking for customers. When such constraints are binding, not only must the plans meet them, but also spillover effects from one period to others occur, according to laws that follow from the theory of individual behaviour under rationing (Samuelson 1947). In particular, consumers willing but unable to borrow are constrained by their current resources, a phenomenon that gives some justification to the Keynesian consumption function relating current consumption to current income.

Neglect of financial constraints may be considered as following from other theoretical simplifications, lack of uncertainty and full knowledge of discounted prices, which rule out insolvency; but it is often particularly restrictive. The role of financial constraints on investment behaviour indeed play a major part in the development of trade cycle theories (Haberler 1937).

Another notable feature of the theory is the simplicity of the trading relations that it assumes. Consumers and producers buy from ‘the market’ or sell to ‘the market’. A worker need not establish ties with a particular employer, nor a manufacturing firm to a particular supplier of raw material. Actually, intertemporal decisions are often subject to quite significant irreversibilities. Long-term commitments are frequent for easily understandable reasons, some of which have to do with the

specificities that characterize many production processes (for instance, most equipment, once bought, cannot be resold). Long-term contracts are also predominant on the labour market, even though many of their clauses often remain implicit. This feature motivates significant research nowadays, under the heading of 'implicit contracts' (Rosen 1985).

Limited as it is, the classical theory of individual intertemporal decisions is, however, indispensable as a starting point, from which the study of the many complexities of real life can proceed. It has moreover brought to light some quite relevant results, such as the fact that, contrary to common belief, the saving of a household need not be an increasing function of interest rates or that individual choice is bound to exhibit some degree of impatience (Koopmans 1960).

An Intertemporal Economy

The theory of general intertemporal equilibrium can also transpose the more familiar static theory. But clearly when so doing it does not go very far; new complications, specific to intertemporal problems, must be faced.

The simple transposition of the general competitive equilibrium assumes the existence of a terminal date, 'the horizon', a given set of consumers and producers whose activities end at this date, if not before. They all decide their plans at the initial date, on the basis of a full system of discounted prices, and acting as price takers. Perfect competition is assumed to imply that discounted prices are such that all markets clear; more precisely for a given date and a given commodity, aggregate supply and demand are defined by addition of corresponding individual supplies and demands contained in individual plans, which may then be considered as fully announced; at equilibrium the aggregate supply is precisely equal to aggregate demand, and this applies for any date and commodity. Hence, all individual plans are, from the initial date, mutually consistent for all future dates.

The usefulness of such an abstract equilibrium concept cannot be judged independently of its application, in particularly for the discussion of properties linking discounted prices to the agents' individual characteristics. Before facing this discussion, it is enlightening to consider how the model can be revised; this was done in three ways.

First, the hypothesis of a full system of markets, one for each date and commodity, has been relaxed and the notion of a *temporary equilibrium* made explicit (Hicks 1939; Arrow and Hahn 1971; Grandmont 1977). Markets then exist only for the exchange of commodities at the (initial) present date, as well as for the loans of one numeraire commodity from the present to the next future date. Thus, present prices and the interest rate of the first period are assumed to be determined by the law of supply and demand, individual plans being made mutually consistent for the initial data. But, in deciding their plans, individual agents have to form anticipations about future prices. Nothing guarantees that these anticipations are correct, so that individual plans will be revised with the passage of time, as actual prices are found to differ from what was expected.

Formal properties of this more realistic model will not be discussed here. Cases can be defined in which anticipations are later realized. It is then possible, but not always necessary when the future is unbounded, that the sequence of temporary equilibria coincides with the equilibrium defined from the hypothesis of a full system of markets. Thus, two sources of difficulty can arise: false anticipations and on the other hand instability following from the myopic functioning of the market system (Hahn 1968).

Second, coming back to the case of a full system of markets, one has relaxed the assumption of a finite horizon with a fixed set of agents. The problem of knowing which firms exist has not been considered as specific to the intertemporal models, and has not been discussed thus far in the framework of these models, given that infinitely lived firms have been assumed. But since the initial proposals of Allais (1947) and Samuelson (1958), consumers are more and more assumed to belong to overlapping generations, each generation living only for a finite time. Such a representation of the consumption sector is clearly more appropriate for long-term analysis than the

assumption of a given set of consumers living for ever, but it raises new difficulties (Balasko and Shell 1980–81).

Third, since long-term phenomena are often involved, it has been found natural and convenient to concentrate attention on specifications in which the exogenous conditions of economic activity, such as technology, tastes, size of the population, natural resources, remain the same through time or change in a simple way; for instance, population increasing at a constant rate while technology exhibits constant returns to scale and natural resources are unbounded. Within such specifications one has dealt with the particular case of a stationary equilibrium, or else with equilibria in which production and consumption all increase at the same constant rate, that is, the case of 'proportional growth'. The analytical usefulness of this assumption of stationarity was at the centre of an important debate on the building of the theory of capital during the 1930s (Knight 1935; Hayek 1936). It follows from the simple form that has the price system of a stationary equilibrium: all discounted prices can be computed from the prices of the present commodities using a single interest rate that applies to all future periods of unit duration. 'The interest rate' is then unambiguously defined (Malinvaud 1953).

Any General Law?

A clear formalization of intertemporal equilibrium not only serves to aid progress in the fundamental conceptualization of economic activity (hence indirectly in the rigour of the discussions concerning many particular questions) but should also lead to comparative statics properties, which, dealing with intertemporal equilibria, have also been called 'comparative dynamics properties'. Particular importance has been given to the question of knowing how the interest rate changes from one stationary equilibrium to another when some specific change is being brought to its exogenous determinants.

The study of this question concentrated on a number of conjectures, which turned out to be about as many disappointments for whose who had expected to find rigorous proofs of their general validity. It is now realized that the rate of interest is related in a very complex way to the many exogenous determinants of equilibrium and that changes of relative prices, which are associated with changes of interest, may be responsible for paradoxical effects. A brief survey of this theoretical search, that extended over many years, nevertheless reveals some basic issues.

Does a high preference of individuals for present consumption necessarily imply a high interest rate? The property was often asserted. When first publishing his *Theory of Interest* in 1907, Irving Fisher called it an impatience theory. Only later when he revised the book for the 1930 edition did he add the subtitle 'as determined by impatience to spend income and opportunity to invest it', which recognizes the role of the productivity of investment (Samuelson 1967). Quite significant cases have indeed been found in the overlapping generation model for which changes of impatience leave the interest rate unchanged (Samuelson 1958).

Does a decrease of the rate of interest mean a lengthening of the production process? The positive answer was taken for granted, at least as long as technology was given, by many economists and was at the head of the 'Austrian Theory' as developed mainly by Böhm-Bawerk (1889) and Hayek (1941). Actually, description of the production process was usually organized in such a way as to focus on the conjectured property, this being true also with such non-Austrian authors as Wicksell (1901). Final output, available for consumption at some date, was seen as resulting from a number of well-identified primary inputs made at previous dates and having 'matured' since then. The notion of an average period of production looked natural; an inverse relationship between this period and the rate of interest was expected. However, it turned out that, even restricting attention to the case of one primary input and one final output, one could not prove the relationship unless a special definition was given to the production period and a special phrasing to the property (Hicks 1939, 1973). Generalization to many primary inputs, many final outputs and many interdependent production processes raises the

fundamental difficulty resulting from induced variations in relative prices; it is quite unlikely that a generalized property could be proved (Sargan 1955).

A somewhat similar property was expected with another formalization that seems to be much more appropriate for describing technology in modern industry. The property concerns the choice of techniques and the notion that different techniques should be selected at various stages of development, as relative scarcity of the two main factors, labour and capital, changes and the interest rate moves accordingly. Its formal specification actually requires a particular model. The production possibility set is seen as resulting from combination of a number of elementary processes, each one operating at constant returns to scale, with fixed input–output coefficients, and requiring a time just equal to one period. Specifying further this model and applying it to an economy with one primary factor (labour), n produced goods and no joint production (the 'Samuelson–Leontief technology'), one defines a technique as a selection of n processes, one for the production of each good.

In this model, given any value of the interest rate, one can determine one technique that is fully appropriate for production, no matter what is the consumption basket. It then seemed natural to conjecture that techniques thus appearing as efficient at different interest rates were ordered from the less capitalistic (high interest) to the most capitalistic ones (low interest). However, this conjecture is not generally valid, even in this special model: as the interest rate progressively declines, one may have to switch at some point away from some technique but have to switch back to it at a later point: this is the case of 'reswitching of techniques' (Morishima 1966).

Is the interest rate systematically smaller when, with a given technology, one shifts from a stationary equilibrium to another one using the same labour input but more productive capital? Again, this looked like a natural property to be stated.

Since in a perfect equilibrium with no uncertainty the net rate of profit must be equal to the interest rate, the property was associated with the notion that capital accumulation must depress profit rates.

The property holds in a purely aggregated model with just one produced commodity, used both for consumption and as productive capital (Solow 1956). The significance of this model for a more general situation was at the heart of hot debates in the late 1950s and early 1960s, the main opponents being located in the two academic cities named Cambridge (Robinson 1956; Lutz and Hague 1961). A side issue was whether one could give unambiguous definitions to such aggregate notions as the volume of productive capital and the marginal productivity of capital. Eventually, both counterexamples and formal analysis of the problem showed that the property was not generally valid (Burmeister and Turnovsky 1972).

The significance of these various negative theoretical results should of course not be overstated. While reflecting the basic complexity of the relationship between the full system of discounted prices and its determinants, the results do not prove that 'pathological cases' are often empirically relevant.

Intertemporal Efficiency

In the same way as the classical theory of individual behaviour, the theory of the optimum allocation of resources can be transposed to the intertemporal framework. Pareto efficiency of a 'programme' made of a set of individual plans, also called 'Pareto optimality', is generalized in an obvious way that need not be spelled out. The two classical duality theorems directly apply as long as the horizon is bounded: the programme resulting from a competitive equilibrium of the type described above is Pareto efficient if no external effect occurs; conversely, under a convexity or atomicity assumption, to any Pareto efficient programme can be associated a set of discounted prices supporting this programme. Properties of this system of prices are similar to those of the competitive price system.

Interesting new applications of these properties may give insights on the evolution of prices

through time. In particular it is easily found that, if extraction costs are negligible, the discounted efficiency price of an exploited exhaustible resource is the same for all future dates, which means that the undiscounted price increases at a rate equal to the interest rate of the numeraire (Hotelling 1931). When forming decisions on the use of exhaustible resources, one should give as much weight to the distant future as to the present; discounting gives no comfort for such decisions.

Theoretical difficulties, however, occur when the more realistic case of an unbounded horizon is being considered. The most relevant of these difficulties concerns the Pareto efficiency of competitive equilibria; efficiency is still proved to hold if the discounted value of the productive capital that exists at date t decreases to zero when one lets t increase to infinity (Malinvaud 1953); but examples of competitive equilibria that do not fulfil this condition and are not Pareto efficient can be found. Such examples may be characterized as cases of overcapitalization, an excessive capital stock being indefinitely maintained without this ever benefiting consumption.

When attention is limited to stationary equilibria, a negative interest rate reveals lack of efficiency, whereas a positive one implies efficiency (if no external effect exists). Similarly, the interest rate of the price system supporting an efficient proportional growth programme cannot be smaller than the rate of growth (Starrett 1970). The borderline case of an interest rate equal to the growth rate corresponds to what was called 'the golden rule'. More precisely, a new notion of optimality has been defined as follows for proportional growth programmes: an optimal programme is feasible and no other feasible programme leads to larger consumptions (that is, a larger consumption of some commodity at some date and no smaller consumption of any commodity at any date). This definition neglects the conditions at the initial date since an 'optimal' programme can require a large input of capital at this date, a larger than is required by other Pareto efficient proportional growth programmes. It was proved that a price system exists that supports such an optimal programme and contains an interest rate equal to the rate of growth (Desrousseaux 1961; Phelps 1961). This is another case in which discounting does not make the distant future negligible.

When it is considered in the preceding terms, the theory of intertemporal efficiency has a somewhat unrealistic aspect; or rather it seems to be quite partial in its treatment of the various questions that intertemporal efficiency raises both for planning and for the study of actual economic evolution. Indeed, the restrictions mentioned in the first section of this article are often serious.

For the theory of planning, even restricted to the medium and long terms, for which intertemporal choices are particularly important, problems concerning the gathering and exchange of information should not be neglected. If a system of discounted prices is to be used for supporting consistency of individual decisions with national objectives, its determination must be given very serious consideration. Moreover, planning often aims at correcting handicaps, distortions or market failures preventing economic development. Its long-term achievement then depends on how well it deals with problems that are not considered here but have motivated an important literature, dealing in particular with the determination of the best shadow discount rate to be used in project evaluation (Dasgupta et al. 1972).

Similarly, for assessing the performance of actual economic systems, one has still to face many questions that again often relate to problems of information. Three of them seem to deserve particular attention. First, the vision of agents exchanging in markets abstracts too much from the complexities of actual contractual arrangements, some of which deal precisely with intertemporal choices; one does not yet clearly see how these complexities react on the behaviour of the full economy, nor even how theory could approach the issue.

Second, the notion of an intertemporal competitive equilibrium should be replaced by that of a sequence of competitive temporary equilibria. It is then known that, even if anticipations are self-fulfilling along this sequence, intertemporal efficiency is not guaranteed; more precisely, the

short-sightedness of equilibria seems to increase the likelihood of an overcapitalization of the type exhibited by the theory of the golden rule. This may occur because of too high saving propensities, because of risk aversion or because of oligopolistic market structures (Malinvaud 1981). But the question of knowing whether and when this likelihood will materialize remains obscure.

Third, the dual assumption of permanent market clearing and permanently equilibrating prices rules out of consideration many issues, such as those arising from variations in the degree of unemployment or in the stimulus given by profitability. A rather common view among supporters of the market system sees these variations as negligible from a long-term perspective, economic evolution being supposed simply to oscillate around the long-term path determined by equilibrium analysis. But critics of the market system and some other economists have the opposite view: economic disequilibria would provide the main clue for an understanding of the comparative growth of nations (Schumpeter 1934; Beckerman 1966). Theory remains conspicuously weak with respect to solving this major debate.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ General Equilibrium
- ▶ Ramsey Model
- ▶ Sequence Economies

Bibliography

Allais, M. 1947. *Economie et intérêt*. Paris: Imprimerie Nationale.

Arrow, K., and F. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Balasko, Y., and K. Shell. 1980–81. The overlapping generations model. *Journal of Economic Theory.* I. The case of pure exchange without money 23(3) (1980): 281–306; II. The case of pure exchange with money 24(1) (1981): 112–142.

Beckerman, W. 1966. The determinants of economic growth. In *Economic growth in Britain*, ed. P.D. Henderson. London: Weidenfeld & Nicolson.

Burmeister, E., and S.J. Turnovsky. 1972. Capital deepening response in an economy with heterogeneous capital goods. *American Economic Review* 62: 842–853.

Dasgupta, P., S. Marglin, and A. Sen. 1972. *Guidelines for project evaluation*. New York: UNIDO, United Nations.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. New York: Wiley.

Desrousseaux, J. 1961. Expansion stable et taux d'intérêt optimal. *Annales des Mines,* November, 31 and 46. Paris.

Drèze, J. 1982. Decision criteria for business firms. In *Current developments in the interface: Economics, econometrics, mathematics*, ed. M. Hazewinkel and A. Rinney Khan. Dordrecht: D. Reidel.

Fisher, I. 1907. *The rate of interest.* New York: Macmillan. 2nd ed., 1930.

Grandmont, J.-M. 1977. Temporary general equilibrium theory. *Econometrica* 45: 535–572.

Haberler, G. 1937. *Prosperity and depression.* Geneva: League of Nations. 3rd enlarged ed., 1941.

Hahn, F. 1968. On warranted growth paths. *Review of Economic Studies* 35: 175–184.

Hicks, J. 1939. *Value and capital*. Oxford: Clarendon.

Hicks, J. 1973. *Capital and time*. Oxford: Clarendon.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.

Knight, F. 1935. The theory of investment once more: Mr. Boulding and the Austrians. *Quarterly Journal of Economics* 50: 36–67.

Koopmans, T.C. 1960. Stationary ordinal utility and impatience. *Econometrica* 28: 287–309.

Lutz, F., and D. Hague, eds. 1961. *The theory of capital*. London: Macmillan.

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

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

Malinvaud, E. 1981. *Théorie macroéconomique*. Vol. 1. Paris: Dunod.

Modigliani, F., and Brumberg, R. 1954. Utility analysis and the consumption function: An interpretation of cross-section data. In *Post-Keynesian economics*, ed. K. Kurihara. New Brunswick/London: Rutgers University Press/George Allen & Unwin, 1955.

Morishima, M. 1966. Refutation of the nonswitching theorem. *Quarterly Journal of Economics* 80: 520–525.

Nickell, S. 1978. *The investment decisions of firms*. Cambridge: Cambridge University Press.

Phelps, E. 1961. The golden rule of capital accumulation. *American Economic Review* 51: 638–642.

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

Rosen, S. 1985. Implicit contracts. *Journal of Economic Literature* 23: 1144–1175.

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

Samuelson, P. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482.

Samuelson, P. 1967. Irving Fisher and the theory of capital. In *Ten economic studies in the tradition of Irving*

Fisher, ed. W.J. Fellner et al. New York: John Wiley & Sons.
Sargan, J.D. 1955. The period of production. *Econometrica* 23: 151–165.
Schumpeter, J. 1934. *The theory of economic development*. Cambridge, MA: Harvard University Press.
Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Starrett, D. 1970. The efficiency of competitive programmes. *Econometrica* 38: 704–711.
von Böhm-Bawerk, E. 1889. *Positive Theories des Kapitales*. Trans. as vol. 2 of Capital and interest. South Holland: Libertarian Press, 1959.
von Hayek, F.A. 1936. The mythology of capital. *Quarterly Journal of Economics* 50: 199–228.
von Hayek, F.A. 1941. *The pure theory of capital*. London: Routledge & Kegan Paul.
Wicksell, K. 1901. *Vorlesungen über Nationalökonomie.* Trans. as *Lectures on political economy.* London: Routledge and Kegan Paul, 1934, vol. 1.

Intertemporal Portfolio Theory and Asset Pricing

Douglas T. Breeden

The intent of this entry is to present intertemporal portfolio theory and asset pricing models, to explain their results and to illustrate the differences between multiperiod and single-period models. To appreciate intertemporal portfolio theory and asset pricing, it is necessary to understand the state of finance theory prior to the seminal intertemporal works of Merton (1969, 1971, 1973), Samuelson (1969), Fama (1970), Hakansson (1970) and Rubinstein (1974). Section "Single-Period Portfolio Theory and Asset Pricing" presents single-period theory and some general results on portfolio statistics. Section "Intertemporal Portfolio Theory" presents intertemporal portfolio theory. Section "Intertemporal Capital Asset Pricing Model (ICAPM)" presents the intertemporal asset pricing model, and Section "Consumption-Oriented Asset Pricing Model (CCAPM)" presents the consumption-oriented representation of it. Section "Extensions and Conclusions" gives important extensions (without proof) and concludes the entry.

Single-Period Portfolio Theory and Asset Pricing

Portfolio choice in terms of means and variances of alternative portfolios' returns was rigorously modelled first in a single-period world by Markowitz (1952, 1959) and Tobin (1958). This theory was significantly extended by Sharpe (1964) and Lintner (1965). By requiring markets to clear in equilibrium, Sharpe and Lintner developed the well-known theory of equilibrium asset prices known as the capital asset pricing model (CAPM). This model was the premier general theoretical model of asset pricing, prior to Merton's (1973) development of the *intertemporal* capital asset pricing model (ICAPM). In fact, despite the development of the theoretically superior (more general) intertemporal asset pricing models, the single-period CAPM is widely used by investment practitioners today.

Portfolio Statistics

In deriving both the single-period and the intertemporal CAPM, there are a few well-known facts about portfolio statistics that are used repeatedly to expedite the derivations. Those will be presented with the notational definitions that follow. First, let $\mathbf{w}^k$ be individual k's $A \times 1$ vector of portfolio weights for risky assets; the ith element represents the fraction of total wealth that is invested in the ith risky asset. From the investor's budget constraint, the amount placed in the riskless asset must be the residual fraction, i.e.,

$$w_0^k = 1 - \sum_i w_i^k.$$

The riskless asset's return is denoted r_f, and risky assets have normally distributed returns with an $A \times 1$ vector of means, μ, and a variance-covariance matrix $\mathbf{V}$. Two statistical results permit the mean and variance of any portfolio and the covariance between any two portfolios' returns to be found from the weights of the portfolios and from the joint distribution of individual assets' returns. (The reader may verify these results from elementary statistical theory on the mean

and variance of a linear combination of random variables.)

Mean portfolio return

$$= \mu_p = w_0 r_f + \mathbf{w}'\boldsymbol{\mu} = r_f + \mathbf{w}'\left(\boldsymbol{\mu} - r_f \mathbf{1}\right). \quad (1)$$

Covariance of 2 portfolios' returns

$$= \sum_i \sum_j w_i^x w_j^y \sigma_{ij} = \mathbf{w}'_x \mathbf{V} \mathbf{w}_y = \sigma_{xy}, \quad (2)$$

where $\mathbf{w}'_x$ and $\mathbf{w}_y$ are the risky asset portfolios and $\mathbf{1}$ is an $A \times 1$ vector of ones. A useful special case of (2) is that the variance of any portfolio's return is $\sigma^2 = \mathbf{w}'\mathbf{V}\mathbf{w}$. Another useful special case of (2) is that, for any portfolio $\mathbf{w}$, the matrix product $\mathbf{V}\mathbf{w}$ gives the $A \times 1$ vector of covariance of all assets returns with the specified portfolio's return. To see this, view each row of the $A \times A$ identity matrix $\mathbf{I}$ as a 1-asset portfolio, and then apply fact (2) row by row to the matrix product $\mathbf{I}\mathbf{V}\mathbf{w} = \mathbf{V}\mathbf{w}$. For reference, these two special cases of (2) will be denoted (2′) and (2″), respectively. Armed with these definitions and facts, we can now expeditiously derive the well-known single-period portfolio theory and CAPM of Sharpe (1964) and Lintner (1965).

Optimal Portfolio Choice

Each individual chooses at time 0 a portfolio that maximizes the expected value of a von Neumann–Morgenstern utility function for wealth at time 1, i.e., $\max E\left[u^k\left(\tilde{W}_1^k\right)\right]$. Since the return on a portfolio is a linear combination of the returns on individual assets, and since the returns on individual assets are assumed to be normally distributed, wealth at time 1 is normally distributed. Thus, given initial wealth W^k, the entire probability distribution for wealth at time 1 is described by the mean and variance of the individual's portfolio return. Rewriting the individual's expected utility as a function of portfolio mean and variance and omitting superscripts for the individual's preferences and portfolio weights, the portfolio choice problem is:

$$\max_{\{w\}} U\left(\mu_w, \sigma^2 w\right) \quad (3)$$

where $\mu_W = r_f + \mathbf{w}'(\boldsymbol{\mu} - r_f \mathbf{1})$ and $\sigma_w^2 = w'\mathbf{V}\mathbf{w}$.

Since individuals like higher mean and lower variance, each portfolio that is maximal for (3) will be 'mean-variance efficient'. Efficient portfolios are those with the highest mean for a given variance, or alternatively, are lowest variance for a given mean.

The choices of the portfolio weights for risky assets are unconstrained in the above problem, since the budget constraint is imposed by making the weight in the riskless asset the residual (negative amounts indicating borrowing). Implicitly differentiating (3) with respect to the vector of risky portfolio weights and setting the partials equal to zero gives a set of linear equations. Solving these by matrix inversion gives the following optimal risky asset portfolio:

$$\mathbf{w}^k W^k = T^k\left[\mathbf{V}^{-1}\left(\mu - r_f 1\right)\right], \quad \text{for all individuals } k, \quad (4)$$

where $T^k = -(\partial U/\partial \mu)W^k/[2(\partial U/\partial \sigma^2)]$ is individual k's compensating variation in variance for a unit change in mean, holding utility constant. Thus, the higher T^k is, the higher k's risk tolerance. Dividing (4) by the sum of the risky asset weights eliminates the individual's wealth and risk tolerance from the new equation, giving the optimum mix of risky asset holdings relative to the total in risky assets.

Thus, we have a remarkable result (first attributed to Tobin 1958): the optimal mix of risky assets in the individual's portfolio depends only upon the means, variances and covariances of risky returns (as perceived by that individual). The individual's current wealth and preferences only affect risky assets' demands through a scalar that is the same for all risky assets. This shows that an individual may separate the choice of the optimal risky portfolio mix from the choice of how much to place in that portfolio and how much in the riskless asset. Sharpe (1964) showed that if all individuals have the same probability beliefs $\{\mu, \mathbf{V}\}$, then the optimal mix of risky assets is the same for all individuals. In fact, if there were a mutual fund that held all risky assets in the proportions given by $\mathbf{V}^{-1}(\mu - r_f \mathbf{1})$, all individuals could achieve their optimal portfolios with that

fund and a riskless asset holding. This property is known as 'two-fund portfolio separation'.

Market Equilibrium. Capital Asset Pricing Model

The aggregate values of individuals' asset holdings, divided by the aggregate market value of wealth of the economy (M), gives 'the market's' portfolio weights. Summing (4) over individuals k and dividing by aggregate wealth M gives the market portfolio. $\mathbf{w}^M$:

$$\mathbf{w}^M = T^M\left[\mathbf{V}^{-1}(\boldsymbol{\mu} - r_f 1)\right], \tag{5}$$

where

$$T^M = \left(\sum_k T^k\right) \Big/ M.$$

Since the market portfolio is a solution to (3) for an appropriate constant, the market portfolio is mean–variance efficient. Pre-multiplying (5) by $\mathbf{V}$ and using the statistical fact (2″), we have that the expected excess returns on assets in equilibrium are proportional to their covariances with the market's return, $\mathbf{V}_{aM}$:

$$\mu - r_f 1 = \left(1/T^M\right)\mathrm{V}_{aM}: \tag{6}$$

Pre-multiplying (5) by $\mathbf{w}'\mathbf{V}$, using formulae for the mean and variance of a portfolio, and rearranging gives the value for the risk tolerance parameter: $(1/T^M) = (\mu_\mathrm{M} - r_f)/\sigma^2{}_M$. The inverse of risk tolerance is termed risk aversion, so higher risk aversion among investors shows up as a higher expected excess return per unit of variance for the market portfolio. Substituting this into (6) gives the well known capital asset pricing model of Sharpe (1964) and Lintner (1965):

$$\text{CAPM}: \quad \boldsymbol{\mu} - r_f 1 = \boldsymbol{\beta}_M\left(\mu_M - r_f\right), \tag{7}$$

where $\boldsymbol{\beta}_M = \mathbf{V}_{aM}/\sigma_M^2$ is the $A \times 1$ vector of assets' betas relative to the market portfolio. They are analogous to the slope coefficients in regressions of assets' returns on the market portfolio's return.

To this date, this single-period capital asset pricing model has been the most widely tested general model of asset prices under uncertainty. It makes the very strong prediction that the expected excess returns across assets are proportional in equilibrium to their betas relative to the market portfolio. Alternatively, it predicts that the market portfolio is mean–variance efficient, in that it gives the highest expected excess return per unit of standard deviation, considering all possible portfolio combinations. Empirical tests of the single-period CAPM usually reject it. Higher beta assets do have higher returns, but the CAPM of (7) is rejected as a representation of the data. Virtually every assumption used in the derivation of the CAPM has been weakened and empirically examined. What follows is the generalization to *multiperiod* or *intertemporal* consumption and investment decisions – probably the most important and productive generalization.

Intertemporal Portfolio Theory

Relaxation of the single-period assumption in portfolio theory has proceeded concurrently in two very similar types of models. First, discrete-time multiperiod models consider individuals who make consumption and investment decisions at fixed points in time, where the interval between decisions is a somewhat arbitrary choice. It is unlikely that an individual would choose only to revise at fixed dates in time, regardless of what happens in between, so these models initially cause concern. However, that concern is alleviated somewhat by the fact that the qualitative properties of optimal policies in many models are unaffected by the choice of updating interval. Key works in discrete-time multiperiod frameworks are those of Samuelson (1969). Hakansson (1970), Fama (1970), Rubinstein (1974, 1976), Long (1974), Dieffenbach (1975), Kraus and Litzenberger (1975), Lucas (1978), Breeden and Litzenberger (1978) and Brennan (1979).

The other model used for intertemporal portfolio theory and asset pricing is the continuous-time model pioneered by Merton (1969, 1971, 1973), and further developed by Cox, Ingersoll and Ross

(1985a, b) and Breeden (1979, 1984, 1986). The continuous-time model assumes that individuals make consumption and portfolio decisions continuously. Although this is not realistic, since individuals do sleep and do things other than make economic decisions, it will not miss important consumption and portfolio adjustments due to the modelling of a fixed time between decisions.

In Merton's continuous-time model, the underlying random processes driving economic uncertainties are assumed to follow continuous-time stochastic processes with normally distributed increments and continuous sample paths. The underlying normality makes the continuous-time model a logical extension for the single-period CAPM and also gives it mathematical tractability that is often not found in discrete-time models. For example, with discrete-time models, a normally distributed stock return results in non-zero probability of a negative stock price. In the continuous-time model, the variance of the stock's return can approach zero as the stock's price approaches zero in such a way as to prevent negative stock prices, but have normally distributed increments at every instant in time. This entry will utilize the continuous-time model, but any important economic intuition found can also be derived in a discrete-time model.

In the intertemporal model, it is assumed that individuals choose consumption and investment policies that maximize their expected utilities across possible *lifetime* consumption paths. In both continuous-time and discrete-time models, preferences are typically assumed to be time-additive and state-independent, i.e., expected lifetime utility for individual k is: $E[\int u^{k}(c^{k}, t)\, dt]$. Although these preferences are not as general as theorists would like, much has been learned with them. It is assumed that the utility of consumption at any instant is monotonically increasing and strictly concave in consumption, in that partial derivatives are: $u_c^k > 0$ and $u_{cc}^k < 0$.

In using the techniques of stochastic dynamic programming to find the best consumption and portfolio policies, it is convenient to break the remaining utility of lifetime consumption into two parts and maximize the sum. At time t the first part is $u^k\ (c^k,\ t)$, the utility of the current consumption over the next period (or instant in time). The second part is the expected utility of consumption for all subsequent periods to that, $J^k(W^k,\ \mathbf{s},\ t)$, which will be explained more fully below. Thus, the objective function is:

$$\max_{\{c,\,\mathbf{w}\}} \left\{u^k\left(c^k, t\right) + E_i\left[J^k\left(W^k, \mathbf{s}, t\right)\right]\right\} \qquad (8)$$

The current choice of consumption affects only the first part directly, but affects the budget constraint for investments made for future consumption. Differentiating (8) with respect to current consumption, taking into account that each additional unit of consumption today is a unit less of investible wealth, gives the standard condition that the marginal utility of consumption equals the marginal utility of wealth for an optimal policy:

$$u_c^k\left[c^k\left(W^k, \mathrm{s}, t\right), t\right] = J_W^k\left(W^k, \mathrm{s}, t\right). \qquad (9)$$

The key difference between single-period portfolio theory and its CAPM and the optimal results in an intertemporal equilibrium arises from the nature of the indirect utility function for wealth. $J(W^k, \mathbf{s}, t)$. The portfolio mix decision affects only the probability distribution of future wealth and therefore only affects J in (8) – the expected utility of future consumption that wealth will be used to buy. The $S \times 1$ vector s is a set of 'state variables' that describe consumption, investment and employment opportunities. When a person expects to live not just for an instant more, but for a period of time, the investment portfolio and consumption rate should be reviewed and adjusted continually. The utility that one expects to get during one's remaining lifetime depends positively on current wealth (since higher wealth buys more goods), but also depends upon the state of investment opportunities. For example, a current wealth of $100,000 provides a lower real consumption stream if the real riskless interest rate is 2 per cent, than if the real rate is 5 per cent. In this case, the real riskless rate is one of the state variables for investment opportunities. Examples of other economic state variables are the expected inflation rate(s) of goods, the

expected productivity of capital or the expected return on the market portfolio, and the level of uncertainty about economic activity or of productivity. Of course, most of these would be considered as endogenous variables; more generally, the underlying exogenous variables could be substituted and the stochastic processes for the endogenous variables derived.

To see the effect of a stochastic investment opportunity set on the investment portfolio, consider a retiree who is relatively averse to risk and holds the single-period optimal portfolio – a little money in the market portfolio and a lot in riskless securities. With that portfolio, the investor has the same wealth if the market is up 10 per cent and the riskless rate is 2 per cent, as when the market is up 10 per cent and the riskless rate is 5 per cent. This may not be optimal, since this retiree has to reinvest his wealth and live off the income. The retiree is financially hurt in the state where the real riskless rate is 2 per cent, and is well off in the 5 per cent state. In addition to the market portfolio, this investor may optimally wish to buy some long-term bonds or interest rate futures contracts that go up in value as rates fall. Then the investor is hedged, by having more wealth to compensate for the poor reinvestment rate. If rates increase, the retiree has a capital loss on the bonds and, therefore, less wealth, but has a better reinvestment rate. Some investors may well prefer this to just holding the market portfolio. Thus, as we shall see, the single-period CAPM's two-fund theorem and the asset pricing model itself will not necessarily hold in a multiperiod economy. These are points all made clear in Merton's (1973) pathbreaking work.

Merton (1973) derived the optimal portfolio rules for an individuals in an exchange economy, and Cox, Ingersoll and Ross (1985b) verified those same portfolio rules in a general equilibrium economy with production. Let subscripts of the indirect utility function be partial derivatives, and let $\mathbf{V}_{aa}$ now be the $A \times A$ variance–covariance matrix for assets' returns and $\mathbf{V}_{as}$ be the $A \times S$ covariance matrix of assets' returns with the various state variables. The optimal portfolio of risky assets in the intertemporal economy is:

$$\mathbf{w}^k W^k = T^k\left[\mathbf{V}_{aa}^{-1}\left(\boldsymbol{\mu} - r_f\mathbf{1}\right)\right] + \mathbf{V}_{aa}^{-1}\mathbf{V}_{as}\mathbf{H}_s^k, \quad \text{for all individuals } k, \tag{10}$$

where $\mathbf{H}^k{}_s = -\mathbf{J}^k{}_{sW}/J^k{}_{WW}$. Notice that the first RHS term of (10) is the mean–variance efficient portfolio as in the single-period equations of (4). As for the other term, Breeden (1979) showed that each column j of the product matrix $V_{aa}^{-1}V_{as}$ represents the portfolio of assets that is most highly correlated in return with movements in state variable j. To see this, note that the portfolio that has the maximum correlation with state variable s_j is the one with the highest covariance with s_j, given a fixed portfolio variance. Mathematically:

$$\begin{aligned}\text{Objective}: \max_{\{\mathbf{w}_j\}} L &= \mathbf{w}_j'\mathbf{V}_{a,sj} + \lambda\left[\sigma^2 - \mathbf{w}_j'\mathbf{V}_{aa}\mathbf{w}_j\right] \\ \text{Solution}: \mathbf{w}_j &= \left[\mathbf{V}_{aa}^{-1}\mathbf{V}_{a,sj}\right](1/2\lambda).\end{aligned} \tag{11}$$

(The scalar does not matter, since all portfolios that are scalar multiples are perfectly correlated and have the same correlations with all other variables.) Thus, those S portfolios are the best hedge portfolios available for individuals to use in hedging opportunity set changes. The coefficient vector in (10), $\mathbf{H}^k{}_s$, gives individual k's holdings of those hedge portfolios (which may be positive or negative).

Aggregating individuals' portfolios gives the market portfolio. Substituting this back into (10) gives:

$$\mathbf{w}^k W^k = \left(T^k/T^M\right)\mathbf{w}^M + \left[\mathbf{V}_{aa}^{-1}\mathbf{V}_{as}\right]\left[\mathbf{H}_s^k - \left(T^k/T^M\right)\mathbf{H}_s^M\right] \quad \text{for all individuals } k, \tag{12}$$

where

$$T^M = \sum_k T^k \quad \text{and} \quad \mathbf{H}_s^M = \sum_k \mathbf{H}_s^k.$$

From this, it is clear that all individuals' portfolios can be obtained with $S + 2$ funds: (1) the market portfolio, (2) the riskless asset, (3) and the S best hedge portfolios for the state variables. No

preferences are needed to set up the mutual funds. Breeden (1984) showed that if each of the S hedge portfolios is *perfectly* correlated with the state variable it hedges, then the allocation of contingent claims is an unconstrained Pareto-optimal allocation (ex ante, as in Arrow 1951). If there is not a perfect hedge for some state variable, then preferences can be chosen so that the allocation is not unconstrained Pareto-optimal.

To complete the analysis, the $\mathbf{H}_S^k$ terms need to be examined, so we know what types of holdings different individuals should have in the hedge portfolios. Without stronger preference assumptions, analysis of the hedging terms is difficult. However, if one assumes that the vector of percentage compensating variations in k's wealth for state variables' changes $\left(\gamma_S^k = -\mathrm{J}_s^k / W^k J_W^k\right)$ are not a function of k's wealth, then Breeden (1984) has shown that:

$$\mathrm{H}_S^k = W^k\left(1 - T^{*k}\right)\gamma_s^k, \tag{13}$$

where T^{*k} is k's Pratt–Arrow measure of relative risk tolerance. Since $\mathbf{H}^k{}_s$ give individual k's holdings of the hedge portfolios for opportunity set changes, an individual will attempt to hedge if and only if his or her relative risk tolerance is less than unity. Since unity represents the logarithmic utility case, those more risk averse than the log will tend to hedge, whereas those more tolerant than the log will tend to 'reverse hedge'. This type of result has been obtained by Merton (1969), Grauer and Litzenberger (1979), Dieffenbach (1975) and Breeden (1984).

The optimality of reverse hedging if relative risk tolerance is greater than unity is a very interesting result, since one certainly cannot rule out those preferences. To understand this result, consider a stochastic expected return on investments in the stock market. Apart from holding the market portfolio, one might wish to hedge or reverse hedge changes in the expected return on the market. For both hedger and reverse hedger, let us assume that an increase in expected return on the market is a good thing, in that expected lifetime utility is positively related to that opportunity. A hedger would say that when the expected return on the market is high, he needs less wealth; on the other hand, when the expected return is low, he needs more wealth to keep up his planned lifetime consumption level.

The person who would reverse hedge would view things differently, but not irrationally. That person would wish to have a lot of wealth to invest when the expected return on the market is high, in order to take advantage of the good returns. When returns are poor, our relatively risk tolerant person would wish to have little wealth to invest. Clearly, this strategy generates a higher multiperiod mean return and a higher multiperiod variance of return than does the hedging strategy. Neither strategy dominates the other for all risk averse individuals. Which is chosen depends upon the person's marginal rate of substitution function of mean for variance.

Intertemporal Capital Asset Pricing Model (ICAPM)

Given the general portfolio theory of the last section, this section derives the general intertemporal asset pricing model of Merton (1973). The first step shows that equilibrium expected returns on all assets are linear combinations of their covariances with the market portfolio and with the S portfolios that are most highly correlated with the opportunity set variables. To see this, aggregate individuals' asset demands (12) to get the market portfolio, premultiply that by $\mathbf{V}_{aa}(M/T^M)$, and rearrange to get:

$$\boldsymbol{\mu} - r_f 1 = [\mathbf{V}_{aM}\mathbf{V}_{as}]\begin{pmatrix} M/T_M \\ -\mathbf{H}_s^M/T_M \end{pmatrix} \tag{14}$$

It is easy to verify that the covariances of assets with the state variables are the same as their covariances with the returns on portfolios that are maximally correlated with the state variables (s^*), which have weights of $\mathbf{w}_{s*} = \mathbf{V}^{-1}{}_{aa}\mathbf{V}_{as}$.

The next step is to derive the expected excess returns on the $S + 1$ mutual funds that individuals hold. Pre-multiplying (14) by the matrix of portfolio weights for the $S + 1$ funds, their expected excess returns are:

$$\begin{pmatrix} \mu_M - r_f \\ \boldsymbol{\mu}_{s*} - r_f \mathbf{1} \end{pmatrix} = \begin{pmatrix} \sigma_M^2 & \mathbf{V}_{Ms*} \\ \mathbf{V}_{s*M} & \mathbf{V}_{s*s*} \end{pmatrix} \begin{pmatrix} M/T_M \\ -\mathbf{H}_s^M/T^M \end{pmatrix}. \tag{15}$$

To see the implications of this, note that if the S hedging portfolios were uncorrelated with the market portfolio and with each other, their expected excess returns would be zero in the single-period CAPM. However, in the intertemporal model, the expected excess return on a hedge portfolio is negatively related to the aggregate hedging demand (opposite if reverse hedging), and proportional to the variance of the hedging portfolio's return. Thus, if individuals in aggregate wish to hedge investment opportunities with a portfolio, they bid up its price and bid down its expected return in equilibrium. As shown earlier, with normal hedging, those state variables with the largest compensating variations in wealth will have the largest hedging demands and will deviate the most from the singleperiod CAPM's return predictions, *ceteris paribus*.

The final step in Merton's intertemporal CAPM is to substitute expected excess returns on the $S + 1$ key portfolios from (15) for preference parameters in (14):

$$\begin{aligned} \boldsymbol{\mu} - r_f \mathbf{1} &= [\mathbf{V}_{aM} \mathbf{V}_{as}] \begin{pmatrix} \sigma_M^2 & \mathbf{V}_{Ms*} \\ \mathbf{V}_{s*M} & \mathbf{V}_{s*s*} \end{pmatrix}^{-1} \begin{pmatrix} \mu_M - r_f \\ \boldsymbol{\mu}_{s*} - r_f \mathbf{1} \end{pmatrix} \\ (\text{ICAPM}) &= \boldsymbol{\beta}_{a, Ms*} \begin{pmatrix} \mu_M - r_f \\ \boldsymbol{\mu}_{s*} - r_f \mathbf{1} \end{pmatrix} \end{aligned} \tag{16}$$

Thus, in the intertemporal economy, betas with respect to the market portfolio are not enough to describe the relevant risk of a security. Its covariances with the investment opportunity set also matter for both pricing and optimal portfolios.

Consumption-Oriented Asset Pricing Model (CCAPM)

Following seminal articles on asset pricing in discrete-time economies by Rubinstein (1976), Lucas (1978) and Breeden and Litzenberger (1978), Breeden (1979) showed that Merton's (1973) multi-beta intertemporal CAPM could be re-expressed with a single risk measure. The result found, which is derived below, is that Merton's multi-beta ICAPM reduces to a market price of risk multiplied by the asset's *consumption-beta*, which is its sensitivity of return to percentage movements in aggregate real consumption. This model is the consumption-based capital asset pricing model (CCAPM).

The optimal rate of current consumption in the continuous-time model is a function of the individual's current wealth and the state vector for investment opportunities, $c^k = c^k(W^k, \mathbf{s}, t)$. In the continuous-time model, the first-order Taylor series approximation is correct for the *stochastic* part of consumption movements. (In contrast, a second-order approximation is required to describe the *expected* change in consumption.) Thus, the stochastic movements in consumption, and the covariances of assets' returns with k's consumption changes $\mathbf{V}_{a,\,ck}$, may be written as follows:

$$\begin{aligned} d\tilde{c}^k &= c_w^k \left(d\tilde{W}^k \right) + c_s^k (\mathrm{d}\tilde{\mathbf{s}}) \\ \mathbf{V}_{a,ck} &= \mathbf{V}_{a,\;\;Wk} c_w^k + \mathbf{V}_{as} \mathbf{c}_s^k. \end{aligned} \tag{17}$$

The risk aversion and hedging preference parameters that determine an individual's asset holdings can be rewritten in terms of an individual's direct utility function for consumption. To see this, implicitly differentiate the envelope condition [Eq. (9), superscript k suppressed]:

$$T = -J_W/J_{WW} = -u_c/(u_{cc} c_W) = T_c/c_W \tag{18}$$

$$\begin{aligned} \mathbf{H}_s &= -J_{sW}|J_{WW} \\ &= -\mathrm{c}_s|c_W, \quad \text{for each individual.} \end{aligned} \tag{19}$$

Substituting these formulae into Merton's optimal asset demands, (10), pre-multiplying them by $(c^k{}_W \mathbf{V}_{aa})$ and using (17) to simplify gives:

$$\mathbf{V}_{a,\;ck} = T_c^k [\boldsymbol{\mu} - r_f \mathbf{1}], \qquad \text{for each individual } k. \tag{20}$$

This shows that each individual holds assets in proportions that result in an optimal consumption rate that covaries with each asset in proportion to its expected excess return. The next step is to

aggregate these individual optimality conditions, which shows that each asset's expected excess return is proportional to its covariance with *aggregate* consumption.

Define the 'consumption beta' for any asset or portfolio j, β_j, to be the covariance of j's return with percentage changes in aggregate consumption, divided by the variance of percentage changes in aggregate consumption. Thus, the consumption beta is the slope in the regression of the asset's return on percentage changes in (real, per capita) consumption. The consumption-oriented CAPM (CCAPM) follows easily from the aggregated version of (20), where the risk tolerance parameter is eliminated by using the expected excess return per unit of consumption beta for *any* portfolio M:

$$\boldsymbol{\mu} - r_f \mathbf{1} = [\boldsymbol{\beta}_c / \beta_{MC}](\mu_M - r_f). \qquad (21)$$

Thus, Breeden (1979) showed that Merton's intertemporal CAPM, which required $S + 1$ betas to determine an asset's systematic risks and equilibrium return, can be collapsed into a consumption oriented CAPM, with only a single beta with respect to consumption. This helps the intuition in determining which types of assets should have equilibrium returns that are substantially different in multiperiod economies than in the single-period world of the original market-oriented CAPM. How much the CCAPM representation helps in the testing of the intertemporal model is the subject of much current debate.

In the intertemporal economy, the market portfolio is no longer mean-variance efficient. The portfolio that has the highest correlation of returns with aggregate real consumption is now mean-variance efficient. To see this, pre-multiply an aggregate version of (20) by V_{aa}^{-1}: the LHS gives the maximum correlation portfolio for consumption, and the RHS shows that it satisfies the mean-variance efficiency property of Eq. (4). The reason is simply that in the intertemporal economy one gets paid to take consumption-related risk, and no other. Any portfolio that is not highest correlation with consumption has wasted risk for no additional return.

Our understanding of these results is greatly enhanced by understanding the relation of asset prices to marginal utilities. Hirshleifer's (1970, ch. 9) presentation of the time-state preference model of Arrow (1964a, b) and Debreu (1959) is used extensively by Fama (1970), Rubinstein (1974, 1976), Long (1974), Hakansson (1977), Lucas (1978), Breeden and Litzenberger (1978), Brennan (1979) and Cox, Ingersoll and Ross (1985b) in asset pricing models that were fundamental precursors to the developments here. They showed that the value and fair price of one more share of an asset is the expected marginal utility of its payoffs. The expected marginal utility of its payoffs depends on the expected sizes of the payoffs, the dates they are received and the covariances of their sizes with the marginal utilities at different dates of a unit of consumption or wealth [see (9)]. Assets that have their highest payoffs when consumption is high (positive consumption betas) are paying the most when least needed, i.e., when the marginal utility of consumption is low; they are less valuable and have higher equilibrium required returns than assets that pay most when consumption is down.

The consumption CAPM follows directly from the marginal utility insights, since with time-additive utility functions, consumption at any date t is a sufficient statistic for marginal utility in an intertemporal economy, since the quality of the investment opportunity set also affects the marginal utility of a unit payoff. The reason that covariance with the market (aggregate wealth) determines risk in the single-period CAPM is that with one period, consumption equals wealth. Since marginal utility is one-to-one with consumption, it is also oneto-one with wealth in the single-period model. In that case, consumption betas and market betas are the same and the CCAPM reduces to the CAPM.

Extensions and Conclusions

Space prohibits the proof of other important results that have been proven or can be proven. For example, Breeden and Litzenberger (1978) showed that if the capital markets allocation is unconstrained Pareto-optimal in this intertemporal economy, then each individual's consumption is monotonically

increasing in aggregate consumption and in every other person's consumption. All individuals' consumption rates should go up and down together, though not necessarily proportionally. Each individual's optimal portfolio is the one that results in the highest correlation of the individual's consumption with aggregate consumption. Breeden (1979) showed: (1) All assets have the same covariances with a portfolio that is maximally correlated with aggregate consumption as with consumption itself. As a result, the CCAPM may be stated and tested in terms of assets' betas with respect to that maximally correlated portfolio. (2) With commodity price uncertainty, the CCAPM holds in terms of expected real returns and the betas of real returns with real, per capita consumption. (3) The CCAPM holds without a riskless asset, by just replacing r_f with the expected return on a portfolio that is uncorrelated with real consumption movements. A similar result was shown earlier by Black (1972) for the single-period CAPM. Finally, Bergman (1985) showed that if preferences are time-multiplicative, rather than time-additive, Merton's intertemporal CAPM still holds, but Breeden's (1979) CCAPM extension does not. Thus, the ICAPM is more general than the CCAPM.

In conclusion, the past three decades have seen important developments in the modelling of consumption and portfolio choices under uncertainty. In my opinion, the intertemporal portfolio theory and asset pricing models presented here are the strongest and most useful theoretical models that we currently have. Finally, I must say that there were many more authors that were important to the development of this area than were described in this entry. The bibliography gives a more complete (but still abridged) listing of papers that serious students should read.

See Also

- Arbitrage Pricing Theory
- Asset Pricing
- Capital Asset Pricing Model
- Finance
- Portfolio Analysis

Bibliography

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman, 507–531. Berkeley: University of California Press.

Arrow, K.J. 1964a. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31: 91–96.

Arrow, K.J. 1964b. The theory of risk aversion. In *Aspects of the theory of risk-bearing*, ed. K.J. Arrow. Helsinki: Yrjö Jahnsson Foundation.

Banz, R.W., and M.H. Miller. 1978. Prices for state-contingent claims: Some estimates and applications. *Journal of Business* 51: 653–672.

Beja, A. 1971. The structure of the cost of capital under uncertainty. *Review of Economic Studies* 38: 359–376.

Bergman, Y.Z. 1985. Time preference and capital asset pricing models. *Journal of Financial Economics* 14: 145–160.

Bhattacharya, S. 1981. Notes on multiperiod valuation and the pricing of options. *Journal of Finance* 36: 163–180.

Black, F. 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45: 444–455.

Black, F., and M.S. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7: 265–296.

Breeden, D.T. 1984. Futures markets and commodity options: Hedging and optimality in incomplete markets. *Journal of Economic Theory* 32: 275–300.

Breeden, D.T. 1986. Consumption, production, inflation, and interest rates: a synthesis. *Journal of Financial Economics* 16: 3–39.

Breeden, D.T., and R.H. Litzenberger. 1978. Prices of state-contingent claims implicit in option prices. *Journal of Business* 51: 621–651.

Brennan, M.J. 1979. The pricing of contingent claims in discrete time models. *Journal of Finance* 34: 53–68.

Constantanides, G.M. 1982. Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *Journal of Business* 55: 253–267.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1985a. A theory of the term structure of interest rates. *Econometrica* 53: 385–407.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1985b. An intertemporal general equilibrium model of asset prices. *Econometrica* 53: 385–407.

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

Dieffenbach, B.C. 1975. A quantitative theory of risk premiums on securities with an application to the term structure of interest rates. *Econometrica* 43: 431–454.

Fama, E.F. 1970. Multiperiod consumption-investment decisions. *American Economic Review* 60: 163–174.

I

Ferson, W.E. 1983. Expected real interest rates and aggregate consumption: Empirical tests. *Journal of Financial and Quantitative Analysis* 18: 477–498.
Fischer, S. 1975. The demand for index bonds. *Journal of Political Economy* 83: 509–534.
Garman, M. 1977. A general theory of asset pricing under diffusion state processes. Working Paper No. 50, Research Program in Finance, University of California, Berkeley.
Grauer, F., and R. Litzenberger. 1979. The pricing of commodity futures contracts, nominal bonds and other risky assets under commodity price uncertainty. *Journal of Finance* 34: 69–83.
Grossman, S.J., and R.J. Shiller. 1982. Consumption correlatedness and risk measurement in economies with non-traded assets and heterogeneous information. *Journal of Financial Economics* 10: 195–210.
Hakansson, N.H. 1970. Optimal investment and consumption strategies under risk for a class of utility functions. *Econometrica* 38(5): 587–607.
Hakansson, N.H. 1977. Efficient paths toward efficient capital markets in large and small countries. In *Financial decision making under uncertainty*, ed. H. Levy and M. Sarnat. New York: Academic.
Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy* 86: 971–987.
Hansen, L.P., and K.J. Singleton. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91: 249–265.
Hirshleifer, J. 1970. *Investment, interest and capital*. Englewood Cliffs: Prentice-Hall.
Huang, C.-F. 1985. Information structure and equilibrium asset prices. *Journal of Economic Theory* 34: 33–71.
Kraus, A., and R.H. Litzenberger. 1975. Market equilibrium in a state preference model with logarithmic utility. *Journal of Finance* 30(5): 1213–1227.
Kydland, F.E., and E.C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.
Lintner, J. 1965. Valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47-(February): 13–37.
Long, J.B. 1974. Stock prices, inflation, and the term structure of interest rates. *Journal of Financial Economics* 2: 131–170.
Lucas, R.E. 1978. Asset prices in an exchange economy. *Econometrica* 46: 14–45.
Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 12: 77–91.
Markowitz, H. 1959. *Portfolio selection: Efficient diversification of investment*. New York: Wiley.
Marsh, T.A., and E.A. Rosenfeld. 1982. Stochastic processes for interest rates and equilibrium bond prices. *Journal of Finance* 38: 635–646.
Merton, R.C. 1969. Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics* 51: 247–257.
Merton, R.C. 1971. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* 3(4): 373–413.
Merton, R.C. 1973. An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.
Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34(October): 768–783.
Pratt, J.W. 1964. Risk aversion in the small and in the large. *Econometrica* 32(1–2): 122–136.
Pye, G. 1972. Lifetime portfolio selection with age dependent risk aversion. In *Mathematical methods in investment and finance*, ed. G. Szego and K. Shell, 49–64. Amsterdam: North-Holland.
Richard, S.F. 1974. Optimal consumption, portfolio and life insurance rules for an uncertain lived individual in a continuous-time model. *Journal of Financial Economics* 2: 187–203.
Roll, R. 1977. A critique of the asset pricing theory's tests. Part I: On past and potential testability of the theory. *Journal of Financial Economics* 4: 129–176.
Ross, S.A. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 3: 343–362.
Rubinstein, M. 1974. A discrete-time synthesis of financial theory. In *Research in finance*, vol. 3, 53–102. Greenwich: JAI Press.
Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics and Management Science* 7: 407–425.
Samuelson, P.A. 1969. Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics* 57(3): 239–246.
Sharpe, W.F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 429–442.
Stulz, R.M. 1981. A model of international asset pricing. *Journal of Financial Economics* 9: 383–406.
Sundaresan, M. 1984. Consumption and equilibrium interest rates in stochastic production economies. *Journal of Finance* 39: 77–92.
Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.

Intrahousehold Welfare

Marcel Fafchamps

Abstract

This article focuses on the allocation of tasks and consumption within the household. We first discuss the role of the household in the production of various self-consumed goods

and services. We then turn to the outcome of bargaining between household members, examining the empirical evidence to date. The last section makes the link between intrahousehold welfare and the matching of spouses in the marriage market.

Keywords

Altruism; Bargaining; Becker, G.; Collective versus unitary models of the household; Comparative advantage; Consumption decisions; Distribution of income and wealth; Division of labour; Economies of size; Household production and public goods; Human capital; Identification; Income-pooling test; Inequality; Inheritance; Intrahousehold welfare; Marriage and divorce; Marriage markets; Paternalism; Preferences; Pre-nuptial agreements; Rotten kid theorem; Sharing rule; Social norms

JEL Classifications

O1

There is a voluminous economic literature on intrahousehold issues. This is hardly surprising given the many critical functions that households fulfil. They are the locus of most consumption decisions and human capital investments. By pooling resources, households generate economies of size and shelter members against unemployment and health shocks. Furthermore the formation and dissolution of households play a crucial role in the long-term distribution of income and wealth. Here we focus on intrahousehold welfare which, as Haddad and Kanbur (1990) have shown, is important to our understanding of inequality in general.

Becker was the first economist to become seriously interested in what happens within the household. Becker's contribution, which is nicely summarized in his *Treatise on the Family* (1981), emphasized three things: the organization of production within the household; the way decisions are made within the household; and the formation of couples. All three have a bearing on intrahousehold welfare.

The Household as a Production Unit

The household is a production and consumption unit, self-providing many services such as food preparation, child care and house chores. In developing countries, households also produce much of their own food and housing and fetch their own fuel and water. Becker (1981) pointed out that the organization of production within the household ought to follow economic principles such as the equalization of marginal returns across activities and the allocation of tasks across household members according to comparative advantage.

These simple observations have far-reaching implications because seemingly small differences between household members can have dramatic consequences. To see why, consider the allocation of wage work and household chores between husband and wife. Assume that the tasks are non-divisible and that the return to education is positive in work outside the home and zero in house chores. It follows that the husband will work outside the home if he is slightly better-educated than his wife. Anticipating this, parents may in turn decide not to invest in daughter education but rather to emphasize learning household chores among girls. This results in a self-fulfilling equilibrium in which women receive less education and are confined to household chores. To the extent that education and independent income affect bargaining within the household, such a traditional division of labour may have dramatic consequences on intrahousehold welfare.

The recent empirical literature has cast some doubt on the efficient organization of production within the household. Using data from West Africa, Udry (1996) showed that households do not equalize returns to labour and organic fertilizer across fields managed by different members. Duflo and Udry (2004) provide similar evidence, showing that household labour resources are not optimally reallocated across activities in response to weather shocks. Fafchamps and Quisumbing (2003) show that comparative advantage alone cannot explain the allocation of tasks within Pakistani households. Their evidence also suggests that most household tasks are easy to learn, contradicting Becker's conjecture that learning-

by-doing locks men and women into specific work patterns. Gender differences in career choices and intrahousehold division of labour may reflect different preferences, possibly shaped by social norms, or result from differences in intrahousehold bargaining.

Intrahousehold Bargaining

Most consumption takes place within households sharing a common budget. (In some societies, such as the coastal region of West Africa, spouses keep separate finances. However, whenever they both contribute to a household public good, they can be regarded as deciding consumption jointly.) Certain consumption goods are rival in the sense that consumption by one precludes consumption by another. Food is an example of a rival good. Other consumption goods – such as a house – are non-rival: they are consumed jointly by the members of the household. In the context of intrahousehold welfare, non-rival goods are usually referred to as (household) public goods.

When choosing how to allocate a limited budget to various rival and non-rival goods, the household takes into account the preferences of its members. Formally, let x_i denote a vector of rival goods consumed by individual i and let X denote household public goods. The household's consumption choices can be represented as the solution to an optimization problem of the form:

$$\max_{\{x_1,\ldots,x_n,X\}} \sum_{i=1}^{N} \omega_i U_i(x_i, X) \text{ subject to } \sum_{i=1}^{N} p x_i + qX = y \tag{1}$$

where ω_i is a welfare weight, N is the number of household members, p and q are prices, and y is income. Consumption choices depend not only on individual preferences $U_i(.)$ but also on welfare weights ω_i: individuals with large ω_i have more weight in the household's decision and hence achieve a higher individual welfare. Understanding intrahousehold welfare thus boils down to understanding the factors that affect ω_i.

In two seminal contributions, Manser and Brown (1980) and McElroy and Horney (1981) model intrahousehold bargaining as depending on threat points: when negotiating over how to allocate consumption expenditures, spouses can threaten to walk away from the couple. How much welfare they can achieve on their own determines how much bargaining power they have within marriage. Intrahousehold welfare is predicted to be determined by rules determining the devolution of assets upon divorce (including alimony, child support and welfare payments).

Lundberg and Pollak (1993) argue that the threat of divorce is too extreme to be credible in most everyday situations. Non-cooperation within marriage is a more realistic threat. In this case, intrahousehold welfare is expected to depend on the financial autonomy of spouses, such as rules determining who receives welfare payments or whether married women have independent access to credit. Lundberg et al. (1997) for instance, show that consumption of women's and children's clothing increased when the UK transferred a substantial child allowance from husbands to wives. McElroy (1990) provides a useful discussion of various factors thought to affect intrahousehold bargaining.

The empirical literature has explored these ideas in terms of 'unitary' versus 'collective' models of the household. A household model is said to be unitary if choices do not depend on bargaining power; otherwise it is collective. A household may be unitary for a variety of reasons, for instance because all decisions are taken by the household head, or because all household members have the same preferences over household consumption $\{x_1, \ldots x_N, X\}$. A simple way of testing the unitary model is the income-pooling test: if welfare weights do not depend on bargaining power, consumption choices should depend only on total income, not on bargaining weights. This yields a simple exclusion test that has been widely applied in the literature, often to identify variables affecting intrahousehold bargaining.

Chiappori (1988) has proposed a way of testing the efficiency of the intrahousehold bargaining process. The basic idea is that the solution to

optimization problem (1) can be written as a two-step process. The household first decides how much to allocate to household public goods X and to the rival expenditures $y_i = px_i$ of each household member, with $pX + \sum_i y_i = y$. Then each member maximizes his or her own utility U_i subject to $px_i = y_i$. Intrahousehold bargaining only affects how total expenditures are shared among members, that is, it affects only the share of rival expenditures that goes to each member. This observation yields testable restrictions on cross-equation parameters in a demand system. This is called the 'sharing rule' approach. Browning et al. (1994), for instance, apply this approach to Canadian couples without children and show that allocation of expenditures on each partner depend on their relative incomes.

Both the sharing rule and the income-pooling tests raise empirical difficulties. One difficulty arises whenever utility is transferable and all household members contribute to a household public good. In this case, Bergstrom (1997) has shown that changes in individual incomes do not affect intrahousehold welfare allocation. The reason is fungibility: reducing the income of a household member simply reduces his or her contribution to the household public good.

Another empirical difficulty is that individual preferences are not directly observable. Hence, in order to identify the effect of bargaining power on household choices, we must assume that different categories of household members have systematically different preferences over joint household consumption. The empirical literature has relied on two types of identification strategies to deal with this issue. The first strategy is to rely on stereotypes, such as 'men prefer alcohol and cigarettes' or 'women care more about children'. This strategy permits identification whenever the stereotype is correct. For instance, it has been shown that, when the bargaining power of the wife increases, the household spends more on child nutrition and schooling (see Bergstrom 1997, and the references cited therein). Based on this evidence, it has been argued that increasing the bargaining power of women is a way to improve child welfare. Such interpretation is a double-edged sword, however. It also reinforces a stereotype that could be used to argue that, since women care for children, it is acceptable for society to relegate them to a reproductive role. What we need is empirical evidence based on actual preferences, not stereotypes.

The second identification strategy is to focus on individual consumption of rival goods such as food or clothing. While this is a better strategy, it also has problems. Browning et al. (1994), for instance, show that households in which the wife earns more spend more on female clothing. They interpret this result as evidence that higher income raises a woman's bargaining power. The problem is that a spouse with a higher income probably occupies a higher job position and needs better clothes to go to work. This may generate a reverse causation between income and clothing expenditures, thereby weakening inference.

Spouses probably derive utility from each other's consumption of rival goods. This point was initially made by Becker (1981), who discusses two possible cases, one in which individuals are altruistic – someone else's *utility* enters their preferences – the other in which they are paternalistic – someone else's *consumption* enters their preferences. An example of paternalistic preferences is when a parent does not want a child to smoke, although the child wishes to. In poor countries, differences in health or nutritional status between spouses have sometimes been interpreted as the result of intrahousehold bargaining (Dercon and Krishnan 2000). Yet it would be quite foolish for even the most despotic and selfish husband to starve his wife to death as she would be of no use to him once gone. Hence, even such a husband would care about his wife's consumption.

An interesting illustration of how altruism can affect intrahousehold welfare is the so-called rotten kid theorem. In this theorem, Becker (1981) imagines a parent who, for altruistic reasons, transfers money to a child. The child can try to capture part of the household income, for instance by refusing to work or by diverting household resources. Becker shows that, as long as the parent decides the size of the transfer after capture has taken place, capture only leads to lower household income and hence to a lower transfer. As a result,

the child chooses not to capture because doing so ultimately reduces his consumption.

The Marriage Market

The marriage market is discussed in a separate entry in this dictionary and need not be discussed in detail here. What is important to realize for our purpose is that, if sufficient commitment mechanisms exist, intrahousehold allocation of welfare can be negotiated up front at the time of marriage. For instance, future spouses may anticipate that a wife who earns an independent income has more say in household decisions. As a result, the groom may insist that the bride will never work before agreeing to marry her. Similarly, if devolution of assets upon divorce affects bargaining power, the newlyweds may sign a pre-nuptial agreement that shapes how assets will be divided.

As first pointed out by Lundberg and Pollak (1993), this observation has deep implications regarding policy intervention. If intrahousehold welfare is entirely decided at the time of marriage, then changing the rules applying to married couples affects only those who are already married. Changes to what happens after marriage (for example, devolution of assets upon divorce) have no long-run effect because, once they have been introduced, they are anticipated in the marriage market.

Provided that this reasoning is empirically correct, the policy implication is that the best policy handle to influence intrahousehold welfare is the marriage market itself. The share of household consumption that women can (implicitly or explicitly) negotiate for themselves depends on the assets they bring to marriage. If this is true, helping women then is best achieved in the long run by improving female education and by changing inheritance rules in their favour.

See Also

- ▶ Becker, Gary S. (born 1930)
- ▶ Collective Models of the Household
- ▶ Household Production and Public Goods
- ▶ Marriage and Divorce
- ▶ Marriage Markets
- ▶ Rotten Kid Theorem

Bibliography

Becker, G. 1981. *A treatise on the family*. Cambridge, MA: Harvard University Press.

Behrman, J. 1997. Intrahousehold distribution and the family. In *Handbook of population and family economics*, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Bergstrom, T. 1997. A survey of theories of the family. In *Handbook of population and family economics*, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Browning, M., F. Bourguignon, P.-A. Chiappori, and V. Lechene. 1994. Income and outcomes: A structural model of intrahousehold allocation. *Journal of Political Economy* 102: 1067–1096.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56: 63–90.

Dercon, S., and P. Krishnan. 2000. In sickness and in health: Risk-sharing within households in rural Ethiopia. *Journal of Political Economy* 108: 688–727.

Duflo, E. and Udry, C. 2004. Intrahousehold resource allocation in Côte d'Ivoire: Social norms, separate accounts and consumption choices. Working Paper No. 10498. Cambridge, MA: NBER.

Fafchamps, M., and A. Quisumbing. 2003. Social roles, human capital, and the intrahousehold division of labor: Evidence from Pakistan. *Oxford Economic Papers* 55: 36–80.

Haddad, L., and R. Kanbur. 1990. How serious is the neglect of intra-household inequality? *Economic Journal* 100: 866–881.

Lundberg, S., and R. Pollak. 1993. Separate spheres bargaining and the marriage market. *Journal of Political Economy* 101: 988–1010.

Lundberg, S., R. Pollak, and T. Wales. 1997. Do husbands and wives pool their resources? Evidence from the United Kingdom Child Benefit. *Journal of Human Resources* 32: 463–480.

Manser, M., and M. Brown. 1980. Marriage and household decision making: A bargaining analysis. *International Economic Review* 21: 31–44.

McElroy, M. 1990. The empirical content of Nash-bargained household behavior. *Journal of Human Resources* 25: 559–583.

McElroy, M., and M. Horney. 1981. Nash-bargained household decisions: Toward a generalization of the theory of demand. *International Economic Review* 22: 333–349.

Udry, C. 1996. Gender, agricultural production and the theory of the household. *Journal of Political Economy* 104: 1010–1046.

Invariable Standard of Value

J. E. Woods

I. In Sections I–III, Chapter I of the *Principles*, Ricardo rejected Smith's labour-commanded theory of value in favour of an embodied-labour theory – for a justification of this, see Sraffa's Introduction to Ricardo (1951) or Garegnani (1984). However, in Sections IV and V, he was forced to modify his theory to take account of the effects of movements in income distribution. Thus, Ricardo had isolated two cases where the value of commodities would change – first, when there was an alteration in the amount of labour required, directly and indirectly, in production; and second, when there was a rise or fall in the value of labour, which operated through unequal capital–labour ratios in the different industries. On empirical grounds, Ricardo argued that the first would dominate the second:

> The greatest effect which could be produced on the relative prices of these goods from a rise of wages, could not exceed 6 or 7 per cent . . .. Not so with that other great cause of the variation in the value of commodities, namely the increase or diminution in the quantity of labour necessary to produce them (Ricardo [1821], 1951, p. 36)

Thus, it was assumed in the remaining chapters of the *Principles* that changes in value were caused by changes in embodied labour.

This left the theory of value in an unsatisfactory state. Central to the Ricardian scheme since 'The Essay on Profits' (Ricardo 1952) had been the rate of profits and its relation to the rate of growth; as a corollary, the determination of the laws determining income distribution was regarded as the principal problem in political economy (Ricardo [1821], 1951, p. 5). Yet, in the study of this problem, Ricardo found that the size of the national income, the quantity of capital and the amount of wages all varied with the distribution of income. Though fixed in physical composition, these variables changed because they were measured in terms of values, themselves functions of the distribution of income. What would be the effect then on the rate of profits, r, of an increase in the real wage rate, w? Would r necessarily decrease or could there be a sufficient rise in the value of net income to accommodate increases in both distributive parameters? It was in an attempt to answer such questions that Ricardo turned to the notion of an invariable standard of value and the associated distinction between relative value and real value.

Having identified the two causes of change in the values of commodities, Ricardo could define the characteristics of a standard measure of value: such a commodity would require a constant quantity of embodied labour in its production and have to be invariant with respect to changes in income distribution. 'Of such a measure, it is impossible to be possessed, because there is no commodity which is not itself exposed to the same variations as the things, the value of which is to be ascertained' (Ricardo [1821], 1951, pp. 43–44). Ricardo was defeated by this problem in *The Principles*, assuming gold to be invariable 'to facilitate the object of this enquiry, although I fully allow that money made of gold is subject to most of the variations of other things' (Ricardo [1821], 1951, p. 46). Nor was he able to achieve progress later, as evidenced by his paper on 'Absolute and Exchangeable Value' (Ricardo 1952). II. For an attempt at a partial solution to Ricardo's problem, we have to turn to the Standard Commodity (§ 23 in Chapter IV of Sraffa (1960) is entitled 'An invariable measure of value'). In *Production of Commodities by Means of Commodities*, Sraffa was 'concerned exclusively with such properties of an economic system as do not depend on changes in the scale of production' (p. v). With a given technique of production, embodied labour cannot change; hence, when investigating the existence of an invariable standard of value, it suffices to consider invariance with respect to income distribution.

Consider a single-product industries, circulating capital model, as in Part I of Sraffa (1960), with price equations given by:

$$p' = wl' + (1 + r)P'A \tag{1a}$$

$$p'(I - A)x = 1 \tag{1b}$$

$p = (p_i)$ is the relative price vector, $l = (l_i)$ the vector of direct labour input coefficients, $A = (a_{ij})$ the matrix of input-output coefficients, w and r the uniform rates of wages and profits respectively, and x is the gross output vector. (1b) states that the actual net output of the economy is the *numéraire*. From (1a):

$$p' = wl'(I - (1 + r)A)^{-1} \tag{2}$$

so that, using (1b):

$$w = 1/l'(I - (1 + r)A)^{-1}(I - A)x \tag{3}$$

Then (2) and (3) imply that:

$$p' = l'(I - (1+r)A)^{-1}/l'(I - (1+r)A)^{-1}(I - A)x \tag{4}$$

i.e., equations (1) can be solved to yield w and p as functions of r in (3) and (4) respectively.

'The key to the movement of relative prices consequent upon a change in the wage rate lies in the inequality of the proportions in which labour and means of production are employed in the various industries' (Sraffa 1960, §15, p. 12). It is a straightforward matter to show that relative prices are invariant with respect to income distribution if and only if there is a uniform value-capital/labour ratio in each industry, itself a manifestation of this underlying (mathematical) condition:

$$l'A = \lambda^*(A)l', \lambda^*(A) > 0 \tag{5}$$

i.e., l is a left characteristic vector of A corresponding to the Frobenius root, $\lambda^*(A)$. (For a proof of this statement, see Pasineti (1977) or Woods (1985).)

Assuming that (5) does not hold, what would happen if, following Sraffa, we suppose that prices are constant as the wage, measured in terms of actual national income, is reduced from one and the rate of profits increased from zero. In those industries with a sufficiently low proportion of labour to means of production ('deficit'-industries), the reduction in wage payments is insufficient to met profit payments. On the other hand, there would be industries with a sufficiently high proportion of labour to means of production ('surplus'-industries) for the proceeds of the wage reduction to exceed profit payments.

> There would be a 'critical proportion' of labour to means of production which marked the watershed between 'deficit' and 'surplus' industries. An industry which employed that particular 'proportion' would show an even balance – the proceeds of the wage reduction would provide exactly what was required for the payment of profits at the general rate (Sraffa 1960, § 17, p. 13).

An industry characterized by that 'proportion' would also exhibit it in its means of production, and in its means of production of means of production, etc. The output of such an industry would consist of the same commodities (combined in the same proportions) as does the aggregate of its own means of production – in other words, such that both product and means of production are quantities of the self-same composite commodity' (Sraffa 1960, §24, p. 19).

Sraffa's Standard Commodity, which possesses that particular 'balancing proportion', is given by the semi-positive vector x^* which solves:

$$Ax^* = \lambda^*(A)x^* \tag{6a}$$

$$l'x^* = 1 \tag{6b}$$

As characteristic vectors are unique only up to scalar multiplication, (6b) is the normalization to fix the size of the Standard System. (The connection between the Standard Commodity and the right Frobenius characteristic vector was first perceived by Newman 1962.) When measured in value terms, the ratio between gross output, x^*, and means of production, Ax^*, is invariant to changes in distribution. Calculating this ration when the wage rate is zero, we obtain:

$$\begin{aligned} p'x^*/p'Ax^* &= p'x^*/\lambda^*(A)p'x^* \\ &= 1/\lambda^*(A) = (1 + R) \end{aligned} \tag{7}$$

where R is Sraffa's 'balancing' ratio, identical to the maximum rate of profits. If A is productive, $\lambda^*(A) < 1$ so that $R > 0$. If, in addition, A is indecomposable (all commodities are basic), $x^* > 0$

and is, in fact, the only semi-positive characteristic vector of A. Thus, existence and uniqueness of the Standard proportions are derived straightforwardly from the Perron-Frobenius Theorem (see Pasineti 1977 or Woods 1978).

Let the Standard Net Product be chosen as numeraire, i.e.

$$p'(I - A)x^* = 1 \tag{8}$$

Then, from the price equations (1a) and (8), and the quantity equations (6a) and (6b), it can be shown that:

$$r = R(1 - w) \tag{9}$$

(see Newman 1962; Pasineti 1977; or Woods 1978). That is, the real wage rate–rate of profits curve for the productive technique described by $\{A; l\}$ *is a* downward-sloping straight line.

It is a straight forward matter to show that (9) holds for a one-commodity model. In a multi-commodity model, prices in general vary with income distribution, as in (4), thereby obscuring the relation between distributive parameters. 'Particular proportions, such as the Standard ones, may give transparency to a system and render visible what was hidden' (Sraffa 1960, §31, p. 23). The use of the Standard System is sufficient, not necessary, for the derivation of a downward-sloping w–r curve, for it can be shown from (3) that $dw|dr < 0$ in terms of any *numéraire*. 'It follows that if the wage is cut in terms of *any* commodity . . . the rate of profits will rise; and vice versa for an increase of the wage' (Sraffa 1960, §49, p. 40).

Thus, the Standard System always exists for the single-product industries, circulating capital model, implying a particularly simple form of the w–r curve.

There are three concluding points to be made.

1. The Standard System does not necessarily exist for the pure joint production model of Part II of Sraffa (1960). For such a model, with input and output matrices A and B, the Standard Commodity would have to satisfy:

$$Bx^* = (1 + R)Ax^*. \tag{10}$$

Sraffa considered the possibility that the Standard Commodity contains negative components. Manara (1980) has demonstrated, using simple numerical examples, that a productive multiple-product industries system does not necessarily have a maximum rate of profit; that is, the solution R,x^* of (10) may be negative or complex.

A reformulation of the system in terms of inequalities would give a simple von Neumann model; the van Neumann solution, which exhibits balanced growth, can be thought of as a generalization of the Standard proportions.

2. A positive Standard System does exist for the single-product industries, fixed capital model in chapter X of Sraffa (1960). Furthermore, as demonstrated in Woods (1984), the Standard Commodity can be used to resolve the particular question of choice of technique which arises in this model – namely, the determination of the optimal economic lifetime of machinery.
3. In any analysis of price and wage variation, some commodity must be chosen as *numéraire*. In chapter III, Sraffa (1960) operated with actual net output, as in (1b). When constructing the Standard Commodity, Sraffa supposes 'that there was an industry which employed labour and means of production in that precise proportion, so that with a wage-reduction, and *on the basis of the initial prices*, it would show an exact balance of wages and profits (Sraffa 1960, §21, p. 16, emphasis added). Clearly, this is the intuitive argument underlying the construction of the Standard Commodity. However, it could not be expressed formally other than in terms of the Standard Commodity.

See Also

▶ Standard Commodity

Biblography

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 36: 291–325.

Manara, C. F. 1980. Sraffa's model for the joint production of commodities by means of commodities. In *Essays on the theory of joint production*, ed. L.L. Pasinetti, 1–15. London: Macmillan.
Newman, P. 1962. Production of commodities by means of commodities. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 98: 58–75.
Pasineti, L.L. 1977. *Lectures on the theory of production*. London: Macmillan.
Ricardo, D. 1821. In *Principles of political economy and taxation. Vol. I of Works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.
Ricardo, D. 1952. Papers and Pamphlets, 1815–1823. In *Vol. IV of Works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Woods, J.E. 1978. *Mathematical economics*. London: Longmans.
Woods, J.E. 1984. Notes on Sraffa's fixed capital model. *Journal of the Australian Mathematical Society, Series B, Applied Mathematics* 26: 200–232.
Woods, J.E. 1985. Notes on relative price invariance. *Giornale degli Economisti e Annali di Economia* 44: 135–152.

Inventories

Louis J. Maccini

Inventories consist of stocks of finished goods, goods-in-process, and raw materials and supplies held by business firms. Interest in inventories among economists stems primarily from the observation that inventory fluctuations are an important feature of business cycles.

This observation was first fully documented by Abramovitz (1950) who observed that, although the level of inventory investment is a small fraction of the level of GNP, changes in inventory investment are a large fraction of changes in GNP, especially in recessions. In particular, he pointed out that in the five business cycles in the US between the two world wars the average decline in inventory investment accounted for 47 per cent of the average decline in GNP during contractions. The post-World War II data for the US are even more dramatic. As reported by Blinder and Holtz-Eakin (1983), the average drop in inventory investment was 68 per cent of the drop in GNP during post-World War II recessions. In expansions, inventory movements account for a smaller fraction of movements in GNP, so that over the cycle as a whole inventory fluctuations are less influential than the data on recessions alone would indicate. Nevertheless, the point remains: inventory movements are a key feature of cyclical fluctuations in aggregate output.

To avoid confusion, it should be stressed that inventories appear to be a propagating mechanism, not a causal force, in business cycles. There is no evidence that exogenous shifts in inventory investment are an underlying cause of fluctuations in GNP.

Microfoundations

To provide a framework for analysing movements in inventories, economists have developed optimiezation models of the firm's behaviour. These borrow heavily from models in operations research, which are designed to guide actual firms in managing their inventory positions. The models of economists have been most fully developed for finished goods inventories held by manufacturers. The prime motive for holding inventories in these models is to serve as a buffer stock, that is, to absorb random fluctuations in demand. Among models of buffer stocks, two approaches may be distinguished: the quadratic criterion–linear constraint approach and the stochastic dynamic programming approach.

The former approach was developed in the pioneering book of Holt et al. (1960). There and in other uses of the theory in economics, for example, Belsley (1969), the firm was assumed to choose levels of output and inventories to minimize expected costs subject to an accumulation equation for inventories and an exogenous stochastic process for sales. To bring the price decision into the theory, Hay (1970), Blinder (1982), and others have assumed that the firm possesses monopoly power and maximizes discounted expected profits.

The basic model presumes that the firm sets its price and output before the random component of demand is revealed, and inventories are used to absorb any shocks to demand. Let P_t be price, Q_t output, N_t real sales or demand, H_t, the stock of inventories at the end of the period, and $\in_t$ a non-negative random variable. The firm is then assumed to maximize

$$E_0 \sum_{t=0}^{\infty} \rho^t \left[P_t N_t - c_0 - c_1 Q_t - c_2 Q_1^2 - v\left(H_t - H_t'\right)^2 \right] \tag{1}$$

subject to

$$H_t = H_{t-1} + Q_t - N_t$$
$$N_t = m_0 - m_1 P_1 + \varepsilon_1$$

where ρ is a discount factor, E_0 is an expectation operator, and C_0, C_1, C_2, m_0 are positive parameters.

The motive for the firm to hold inventories is essentially contained in the inventory-holding cost function, $v\left(H_t - H_t'\right)^2$, which is a U-shaped Function of H_t that reaches a minimum at H_t'. It captures two forces: Higher inventories increase costs in the form of storage costs. But, lower inventories also increase costs in the form of lost sales, since lower inventories relative to sales increase the likelihood the firm will be caught out of stock. The holding cost function balances these forces.

The model has several advantages. It yields explicit solutions for the choice variables that are linear in the previous period's inventory stock and current and future sales. Further, certainty equivalence applies so that random variables may be replaced by their expected values. But there are disadvantages. The functional form assumptions are quite strong. The random component to demand must appear additively. And the discount rate must be constant, which limits the ability of the model to analyse the effects of changes in real interest rates through which monetary policy operates.

The stochastic dynamic programming approach to inventory-holding behaviour originates with the work of Arrow et al. (1951), who also developed a model of a firm that minimizes expected costs subject to an exogenous stochastic process for sales. Karlin and Carr (1962) later extended the model to allow for a price decision. The economic implications of the model have been drawn out by, among others, Mills (1962) and Zabel (1972).

Suppose the firm makes decisions over a planning horizon of T periods. Then, according to Bellman's Principle of Optimality, and using the earlier notation, the functional equation that describes the firm's optimal programme is

$$\begin{aligned} \Omega_T(H_{t-1}) = \max_{P_t, Q_t} \{ & P_t m(P_t) - P_t D(b_t) - c(Q_t) \\ & - A(b_t) + \rho_t \int_0^{\infty} \Omega_{T-1}(H_t) f(\varepsilon_t) \mathrm{d}\varepsilon_t \Big\} \end{aligned} \tag{2}$$

where

$$\begin{aligned} b_t &= H_{t-1} + Q_t - m(P_t) \\ H_t &= b_t - \varepsilon_t \\ D(b_t) &= \int_{b_t}^{\infty} (\varepsilon_t - b_t) f(\varepsilon_t) \mathrm{d}\varepsilon_t \\ A(b_t) &= \int_0^{b_t} a(b_t - \varepsilon_t) f(\varepsilon_t) \mathrm{d}\varepsilon_t \end{aligned}$$

and where $m' \leq 0$, $m'' \leq 0$, $c' > 0$, $c'' \geq 0$ $a' > 0$, and $a'' \geq 0$. The expression, $\Omega_{T-1}(H_{t-1})$, denotes maximum discounted expected profits over t future periods.

Observe that $D(b_t)$ is expected shortages or stockouts, capturing the real sales that the firm can expect to lose if it is caught with too few inventories, while $A(b_t)$ is the expected storage cost of holding inventories. Unlike the quadratic cost-linear constraint approach, this model permits general functional forms to be used. Moreover, the decision rules will depend in general on moments higher than the mean of the probability distribution of demand. Finally, although the model has been formulated above with additive demand errors, it can be modified to allow for non-additive uncertainty. A disadvantage of the model is that explicit solutions are not possible.

In both approaches, the incentive for the firm to hold inventories is essentially the same. To see this, observe that expected shortages, $D(b_t)$, in (2) are inversely related to initial inventories

and correspond to the cost savings to holding inventories – the downward-sloping section of the U-shaped cost curve – that appears in (1). In essence, both terms capture the benefits to the firm of holding a buffer stock in the form of finished goods inventories. Given this, it is not surprising that with additive demand uncertainty both approaches yield similar economic predictions, though of course the details will differ from model to model. Such predictions have served as the basis for much empirical work on inventories and have motivated the specification of inventory investment relationships in macroeconomic models.

One set of predictions concerns the response of decision variables to changes in initial inventories. As long as marginal production costs rise (i.e., $c_2 > 0$ or $c'' > 0$) so that there is some incentive to smooth production, both P_t and Q_t will be inversely related to $H_{t\text{-}1}$. Further, $-1 < (\partial H_1/\partial H_{t-1}) < 0$. This means that inventory adjustments will exhibit the characteristics of partial adjustment.

A second set of predictions refers to the effects of changes in exogenous variables. Generally speaking, P_t and Q_t will be positively related to exogenous shifts in anticipated demand (shifts in m_0 or the function $m(P_t)$) and inversely related to changes in inventory holding costs. Further, increases in current anticipated demand, declines in future anticipated demand, or increases in inventory holding costs will reduce inventory investment.

The basic models outlined above have been extended in numerous directions. A prominent one is that right from the beginning many authors have allowed production costs to depend on the change in the level of output as well as the level of output itself. This creates an additional incentive for the firm to smooth production beyond the incentive embodied in the assumption that $c'' > 0$. The difficulty with this notion is that the underlying rationale for including costs to changing the level of output is quite vague. A possible rationale that has been exploited in recent work – for example, Maccini (1984) – is that such costs reflect adjustment costs to changing quasi-fixed factors of production, for example, plant and equipment, or the stock of workers. But, then, a better theoretical procedure would be to incorporate decisions on quasi-fixed factors directly into the analysis. This more explicit specification of the economic forces at work permits an analysis of the important interaction between inventories and quasi-fixed factors of production.

Most inventory models that analyse price as well as output and inventory behaviour, have assumed that the firm possesses some monopoly power in output markets. This assumption facilitates a study of the relationship between price and inventory behaviour, but it is not essential. Models of a competitive industry can be formulated to yield similar predictions for movements in price, output and inventories – see, for example, Eichenbaum (1983).

So far we have concentrated, like the literature, on the holding of finished goods inventories. Theory that rationalizes the holding of inventories of goods-in-process and raw materials and supplies is much less well developed. It is common to adapt, rather casually, the earlier models designed for the study of finished goods inventories. But, this tactic fails to capture in a rigorous way the firm's motives in holding these other inventories, and may in fact generate specious predictions. This is an area that needs more research.

Finally, all the theories discussed so far are designed to explain the holding of manufacturers' inventories, and are not obviously applicable to the holding of inventories by wholesalers and retailers. The trouble is that for the latter agents the variable corresponding to Q_t – the variable that gives rise to additions to inventories – is orders of goods from manufacturers. But, the process of ordering and receiving goods from manufacturers may carry a substantial fixed cost which will induce the firm to 'bunch' rather than to 'smooth' orders and deliveries. In recent work, Blinder (1981) has used ideas associated with S-s models of inventory behaviour – see, for example, Scarf (1960) – together with an aggregation procedure to undertake an analysis of retail inventories. This is an important development because retail inventories appear to exhibit at least as much volatility as manufacturers' inventories.

Empirical Work

Empirical studies of inventories have been dominated by the use of the flexible accelerator model, first used in empirical work by Lovell (1961), and subsequently used by many authors, including Blinder and Holtz-Eakin (1983), Maccini and Rossana (1984), Irvine (1981), etc. The model takes the form

$$H_t - H_{t-1} = \lambda_1 \left(H_t^* - H_{t-1}\right) + \lambda_2 [N_t - E_t(N_t)] + u_i, \quad 0 \le \lambda_1 \le 1, \quad -1 \le \lambda 2 \le 0 \tag{3}$$

where H_t^* is 'desired' inventories, $E_t(N_t)$ is expected sales so that $N_t - E_t(N_t)$ is a 'sales surprise', and u_t captures random forces other than sales that operate on inventories. Desired inventories should depend on current and future levels of expected demand, inventory holding costs, such as real interest rates, and factor input prices which shift marginal production costs.

This model embodies the essential ideas of the buffer stock models of inventory behaviour described above. As long as marginal production costs rise with output, $c_2 > 0$ or $c'' > 0$, so that there is an incentive for firms to smooth production, it follows that $\lambda_1 < 1$ so that firms will partially adjust inventory stocks to desired levels and that $\lambda_2 < 0$ so that sales surprises will be met partly by inventory adjustments. In the extreme, as $c_2 \to \infty$ or $c'' \to \infty$, it follows that $\lambda_1 \to 0$ and $\lambda_2 \to -1$, and vice versa. Moreover, the determinants of desired inventory stocks reflect the main exogenous variables that are embedded in the models of the firm's behaviour.

This model has been estimated over a wide variety of industries, sectors, countries, and sample periods. Different assumptions have been used for expectation formation schemes, and different econometric techniques have been used to handle statistical problems.

Despite the enormous amount of empirical work done with the model, a number of empirical puzzles remain. A major one surrunds the estimates of the parameters, λ_1 and λ_2, as several authors, most prominently Feldstein and Auerbach (1976) have pointed out. The estimates of λ_1 turn out to be very low, implying that the speed of adjustment of actual to desired inventories is very slow. This is implausible when wide savings in inventory investment amount to no more than a couple of weeks of production. Moreover, the estimates of λ_2 are also quite low, implying that sales surprises tend to be absorbed largely by production adjustments. But, this contradicts the low estimates of λ_1, which suggest that it is difficult, that is, costly, to adjust production levels in order to close gaps between H_t^* and H_{t-1}. Despite the use of relatively sophisticated econometric methods in recent work to deal with statistical problems, the puzzle remains unresolved.

Another puzzle is that real interest rates have generally performed poorly in inventory equations – see, for example, Blinder and Holtz-Eakin (1983) and Maccini and Rossana (1984) for recent studies. This is a surprise since it is widely believed that changes in credit conditions have a substantial effect on the inventory positions of business firms. This result may be owing to difficulties in measuring *ex ante* rates, or to a lack of variation in real rates, or to the presence of credit rationing which is not adequately captured in the price of credit.

Finally, a number of authors (e.g. Blinder and Holtz-Eakin 1983), have observed that the variance of production actually exceeds the variance of sales in most industries. This fact appears to conflict with the above theoretical models which predict that inventories are used to smooth production relative to sales. To explain this phenomenon, the models need to be extended to allow for cost shocks in the form of, for example, real raw material prices, or more complex demand structures than serially uncorrelated random errors.

See Also

- ▶ Adaptive Expectations
- ▶ Adjustment Costs
- ▶ Buffer Stocks
- ▶ Cobweb Theorem
- ▶ Inventory Cycles
- ▶ Inventory Policy Under Certainty
- ▶ Investment (Neoclassical)

► Layoffs
► Stochastic Optimal Control

Bibliography

Abramovitz, M. 1950. *Inventories and business cycles*. New York: National Bureau of Economic Research.
Arrow, K., T.B. Harris, and J. Marschak. 1951. Optimal inventory policy. *Econometrica* 19: 250–272.
Belsley, D. 1969. *Industry production behavior: The order stock distinction*. Amsterdam: North-Holland.
Blinder, A. 1981. Retail inventory behavior and business fluctuations. Brookings Papers on Economic Activity No. 2,261G21 443–505.
Blinder, A. 1982. Inventories and sticky prices: More on the microfoundations of macroeconomics. *American Economic Review* 72(3): 334–348.
Blinder, A., and D. Holtz-Eakin. 1983. Inventory fluctuations in the United States since 1929. Proceedings of the conference on Business Cycles, National Bureau of Economic Research.
Eichenbaum, M. 1983. A rational expectations equilibrium model of inventories of finished goods and employment. *Journal of Monetary Economics* 12(2): 259–277.
Feldstein, M., and A. Auerbach. 1976. Inventory behaviour in durable goods manufacturing: The target adjustment model. *Brookings Papers on Economic Activity* 2: 351–396.
Hay, G. 1970. Production, price and inventory theory. *American Economic Review* 60(4): 531–545.
Holt, C., F. Modigliani, J.B. Muth, and H. Simon. 1960. *Planning production, inventories and work force*. Englewood Cliffs: Prentice-Hall.
Irvine, O. 1981. Retail inventory investment and the cost of capital. *American Economic Review* 70(4): 633–648.
Karlin, S., and C. Carr. 1962. Prices and optimal inventory. In *Studies in applied probability and management science*, ed. K. Arrow, S. Karlin, and H. Scarf. Palo Alto: Stanford University Press.
Lovell, M. 1961. Manufacturers' inventories, sales expectations and the acceleration principle. *Econometrica* 29: 293–314.
Maccini, L. 1984. The interrelationship between price and output decisions and investment decisions: Microfoundations and aggregate implications. *Journal of Monetary Economics* 13(1): 41–65.
Maccini, L., and R. Rossana. 1984. Joint production, quasi-fixed factors of production, and investment in finished goods inventories. *Journal of Money, Credit, and Banking* 16(2): 218–236.
Mills, E. 1962. *Price, output and inventory policy*. New York: Wiley.
Scarf, H. 1960. The optimality of (S-s) policies in a dynamic inventory problem. In *Mathematical methods in the social sciences*, ed. K. Arrow, S. Karlin, and P. Suppes. Palo Alto: Stanford University Press.
Zabel, E. 1972. Multiperiod monopoly under uncertainty. *Journal of Economic Theory* 5(3): 524–536.

Inventory Cycles

Michael C. Lovell

The Contribution of Inventory Liquidation to Declines in GNP

Although inventory investment is a relatively small component of total GNP, even in boom years, the swing from inventory accumulation to massive liquidation is a fundamental factor in the propagation of cyclical reversals in the pace of economic activity. To illustrate, the 1981–2 decline in United States GNP of $31.8 billion dollars deserves to be called an inventory recession because the shift from positive inventory accumulation of $8.9 billion (1972 dollars) at the cycle peak to liquidation of stocks at an annual rate of $22.7, a decline in effective demand of $31.6 billion, greatly exceeded the collapse of any other component of real GNP.

As inspection of Table 1 makes clear, a drop off in inventory investment generally makes a major contribution to each recession's decline in effective demand. The short 1980 recession should probably not be counted as an exception to this rule in that inventory investment fell from $13.7 billion in the second quarter of 1979 to − $10.1 billion in the third quarter of 1980. However, the 1946 recession (which pre-dates the availability of quarterly GNP data) is atypical; this was not an inventory recession because the efforts of business enterprise to replenish stocks at the end of World War II served to soften the shock of postwar conversion. And in the Great Depression of the 1930s, the decline in inventories was overshadowed by a collapse of fixed investment that helped push the unemployment rate up to 25 per cent.

The critical importance of inventories has long been recognized, thanks in large measure to the fundamental empirical study of Moses Abramovitz (1950), who demonstrated that in

Inventory Cycles, Table 1 Contribution of inventory disinvestment to cyclical declines in GNP

GNP turning point date (Year & quarter)	Gross national product	Inventory investment	Fixed investment nonresidential	Residential investment	Δ Inventory/Δ GNP (%)
Peak: 1948:4	497.9	5.3	51.9	24.1	
Trough: 1949:4	490.8	−7.7	43.5	26.9	
Change	*−7.1*	*−13*	*−8.4*	*2.8*	*183.10*
Peak: 1953:2	628.3	5.1	55.9	28.2	
Trough: 1954:2	608.1	−4.1	54.8	29	
Change	*−20.2*	*−9.2*	*−1.1*	*0.8*	*45.54*
Peak: 1957:3	688.5	3.7	67.3	28.9	
Trough: 1958:1	665.5	−6.8	61.5	28.2	
Change	*−23*	*−10.5*	*−5.8*	*−0.7*	*45.65*
Peak: 1960:1	740.7	12.7	67.4	37.3	
Trough: 1960:4	732.1	−5.3	66.3	32.7	
Change	*−8.6*	*−18*	*−1.1*	*−4.6*	*209.30*
Peak: 1969:3	1092	13.7	118.5	43.2	
Trough: 1970:1	1081.4	2.1	115.4	40.6	
Change	*−10.6*	*−11.6*	*−3.1*	*−2.6*	*109.43*
Peak: 1973:4	1266.1	23.7	140.7	57.4	
Trough: 1975:1	1204.3	−14.3	120.7	39.4	
Change	*−61.8*	*−38*	*−20*	*−18*	*61.49*
Peak: 1980:1	1496.4	−0.5	171.8	53	
Trough: 1980:2	1461.4	−2.1	162.2	42.4	
Change	*−35*	*−1.6*	*−9.6*	*−10.6*	*4.57*
Peak: 1981:2	1512.5	8.9	167.1	47.3	
Trough: 1982:4	1480.7	−22.7	181.3	40.6	
Change	*−31.8*	*−31.6*	*14.2*	*−6.7*	*99.37*
				Average	*94.81*

Note: All GNP magnitudes measured in 1972 dollars.

the period between the two world wars a collapse of inventory investment contributed decisively to each recession's decline in effective demand. Inventories have continued to play a destabilizing role in each United States recession since the publication of Abramovitz's study. Also, substantial empirical evidence for a number of countries establishes that inventory recessions are a general characteristic of capitalist economies. Further, there is some evidence, reviewed by Attila Chikàn (1984), that socialist economies may also experience inventory cycles.

The phrase 'inventory recession' stresses the empirical fact that cyclical reversals in economic expansion are dominated by the liquidation of inventory stocks. But to characterize most cyclical reversals as 'inventory recessions' does not explain why the declines in the pace of economic activity come about or what policy measures, if any, should be applied to mitigate the sacrifice of jobs and output occasioned by recession. Empirical observation, such as the evidence of Table 1, cannot by itself establish that inventories are in any sense a fundamental cause of the business cycle rather than only a basic symptom or but one of a number of essential ingredients of the mechanism propagating business fluctuations. Both theory and empirical evidence are essential in the study of the inventory cycle mechanism.

Modelling the Inventory Cycle

In the interwar period, members of both the psychological and the monetary schools of business cycle theory indicted inventory investment as a

particularly critical factor in the generation of business cycles. R.G. Hawtrey (1928) argued that monetary factors had their primary effect on the economy through their influence on inventory investment. Pigou (1929) also stressed the impact of systematic errors of optimism and pessimism in explaining cyclical movements.

A major contribution to our understanding of inventory cycles was made by Erik Lundberg (1937), who showed how a set of quite simple assumptions suffices to generate cycles in economic activity, as illustrated by the data for a hypothetical inventory cycle reported on Table 2. Observe that a once and for all step increase in autonomous government spending from 500 to 600 in period 5 (reported in column 1) disturbs the initial equilibrium, leading to cycles in output (column 5), sales (column 7), and inventory investment (column 8). Output rises from the initial equilibrium of 1200 to a peak of 1838 in period 8 and then slumps to a recession low of 1233 in period 13. Thus the attempts by business enterprises to use inventories as a buffer to insulate output from sales are frustrated in the aggregate; indeed, production fluctuates *more* than sales volume (compare columns 5 and 7) over the course of the inventory cycle.

The following details of this inventory cycle deserve notice: In each time period the entries are determined in accordance with the simple assumptions enumerated at the bottom of Table 2. Initially inventories serve as a buffer permitting business firms to meet the unanticipated increase in demand occasioned by the increase in government spending. The immediate impact of the increase in government spending is limited to a drawing down of inventories by 100 units in period 5, as reported in columns 8 and 9; there is no immediate change in output because business firms did not anticipate the increase in sales when scheduling production for period 5. For period 6 output of 1500 is scheduled in order to meet anticipated sales of 1350 (same as last period) plus 150 units to be added to inventory in order to both replace the items sold

Inventory Cycles, Table 2 Lundberg–Metzler inventory cycle model

MPC = 0.60					MDIC = 0.50				
Time period	Government spending (1)	Anticipated sales (2)	Planned inventories (3)	Planned change in inventories (4)	Output (5)	Consumption (6)	Actual sales (7)	Actual invention investment (8)	Inventory stock (9)
1	500	1250	625	0	1250	750	1250	0	625
2	500	1250	625	0	1250	750	1250	0	625
3	500	1250	625	0	1250	750	1250	0	625
4	500	1250	625	0	1250	750	1250	0	625
5	600	1250	625	0	1250	750	1350	−100	525
6	600	1350	675	150	1500	900	1500	0	525
7	600	1500	750	225	1725	1035	1635	90	615
8	600	1635	818	203	1838	1103	1703	135	750
9	600	1703	851	101	1804	1082	1682	122	872
10	600	1682	841	−30	1652	991	1591	61	932
11	600	1591	796	−137	1454	873	1473	−18	914
12	600	1473	736	−178	1295	777	1377	−82	832
13	600	1377	688	−144	1233	740	1340	−107	725
14	600	1340	670	−55	1285	771	1371	−86	639
15	600	1371	685	46	1417	850	1450	−33	606
16	600	1450	725	119	1569	942	1542	28	634
17	600	1542	771	137	1679	1007	1607	71	705
18	600	1607	804	98	1706	1023	1623	82	787
19	600	1623	812	24	1648	989	1589	59	846
20	600	1589	794	−52	1536	922	1522	15	861
New equilibrium									
	600	1500	750	0	1500	900	1500	0	750

out of inventories in period 5 and to increase the inventory stock to the higher level of 675 (column 4) which is desired because of increased business volume. But in spite of these adjustments, the economy does not achieve equilibrium in period 6; once again the immediate impact of an excess of sales over anticipations is met by drawing down inventories below the desired level. The boom is temporary, for eventually inventories catch up with the expanding economy. The economy gradually converges in a series of oscillations toward the equilibrium level presented in the bottom row of the table.

While the cycle displayed on Table 2 is stable, the economy converging in the limit to equilibrium, Lloyd Metzler (1941) showed analytically how the stability of the inventory cycle developed by Lundberg depended critically on the two parameters of the model, the Marginal Propensity to Consume (MPC) and the Marginal Desired Inventory Coefficient (MDIC). He proved that the model will converge toward equilibrium rather than explode only if the following stability condition is satisfied:

$$\text{MPC} \times (1 + \text{MDIC}) < 1.0$$

For the parameter values used in constructing Table 2 (MPC = 0.60 and MDIC = 0.5) we have $0.6 \times (1 + 0.5) = 0.9 < 1.0$, as required for stability; but if the reader recalculates the table with the MPC = 0.6 but MDIC = 0.7, a series of divergent cycles will be observed ($0.6 \times (1 + 0.7) = 1.02$). Metzler also considered the implications of replacing the assumption that sales are expected to be the same as last period with extrapolative expectations.

Assumptions of the Lundberg–Metzler Inventory Cycle Model

1. Government spending (column 1) increases from 500 to 600 in period 5, generating cyclical movements in sales, output, consumption and inventory investment.
2. Expectations are "static", for anticipated sales in the current period (column 2) always equal actual sales in the preceding period (carried over from column 7).
3. Planned inventories (column 3) equal the marginal desired inventory coefficient (MDIC) times anticipated sales (column 2).
4. Planned inventory change (column 4) equals the excess of planned inventories over last period's inventory stock (column 9).
5. Output (column 5) is the sum of anticipated sales plus the planned change in + column 4.
6. Consumption (column 6) is the Marginal Propensity to Consume (MPC) times output.
7. Actual Sales (column 7) equal Government Spending plus Consumption (column 1 + column 6).
8. Actual Inventory Investment (column 8) equals output less sales (column 5–column 7).
9. The Inventory Stock (column 9) increases from the preceding period by Actual Inventory Investment (column 8).

While the Lundberg–Metzler model demonstrates how the inventory cycle can result from a few quite simple assumptions, simplicity has its costs: Lundberg and Metzler did not show that their assumptions about firm behaviour were compatible with the assumption of maximizing behaviour, they neglected problems of aggregation involved in moving from assumptions about firm behaviour to macro aggregates, they neglected the influence of monetary factors, and they assumed that the adjustment to surprise is met entirely through shifts in buffer stock inventories rather than through price reductions or adjustments in advertising expenditure. Their analysis did serve to inspire the investigation of these and several other issues by a number of authors. Holt and Modigliani (1961) showed how the Lundberg– Metzler assumptions about output determination could be derived from the assumption of firm optimizing behaviour, where the task faced by the entrepreneur involved the minimization of a dynamic quadratic loss function; of course, they also found that alternative specifications of the loss function would yield an embarrassing wealth of alternative behavioural equations. Lovell (1962) showed that the Lundberg–Metzler behavioural assumptions led to instability for any reasonable set of parameters when the problem of aggregating from the firm to

the economy-wide level was addressed within the context of a multi-sector dynamic input–output model; while stability might still be obtained by replacing the Lundberg–Metzler assumption that firms attempt an immediate one-period correction of inventory imbalances with the flexible accelerator assumption that only a fraction of the gap between desired and actual inventories is eliminated each time period, he also showed that the system was necessarily unstable if firms had perfect expectations, correctly anticipating the volume of sales in the next period – that is to say, systematic expectational errors under certain circumstances contribute to stability. Lovell (1974) also showed how the real inventory cycle of Lundberg–Metzler could be influenced by monetary policy once desired inventories were assumed to depend on the interest rate as well as sales volume. In a more recent study, Blinder and Fisher (1981) find that the introduction of inventories modifies the standard rational expectations macroeconomic model in two important respects: first, cyclical rather than random fluctuations are generated when the economy is disturbed by unanticipated monetary shocks; second, if desired inventories are sensitive to real interest rates, even fully anticipated changes in money can affect real economic variables, which contradicts the policy impossibility result of rational expectations theory. Adding inventories to the simplified rational expectations model leads to an explanation of the cycle and simultaneously offers the hope for mitigating fluctuations through appropriate policy action.

Empirical Research

Recognition of the important role of inventories in recessions places a number of basic questions on the agenda for empirical research:

First, are the behavioural assumptions about firm behaviour made by Lundberg and Metzler consistent with the observations? Beginning in the mid 1950s, a number of studies established that the Lundberg–Metzler model is consistent with the evidence provided response lags are introduced; most investigators found that the flexible accelerator assumption that firms attempt only a partial adjustment of inventories toward their equilibrium level within a single period appears more appropriate than the assumption that firms engage directly in 'production smoothing'.

Second, are expectations of future sales volume subject to substantial error, are they rational, and do they have a substantial impact on inventory holdings? Investigators have in recent years been inclined to adopt the assumption of rational expectations as part of the maintained hypothesis rather than subjecting it to direct empirical test; where the rationality hypothesis has been tested, as in Hirsch and Lovell (1969), it has *not* dominated alternative models, such as the extrapolative model of Robert Ferber or the adaptive expectations model. Expectations may be subject to systematic error and forecast errors do affect inventory holdings; but the mechanism is not as simplistic as Lundberg and Metzler assumed in constructing their inventory cycle models.

Third, how big an impact do changes in nominal interest rates and anticipated price changes have on inventory holdings? Prior to the inflationary era of the 1970s, investigators were usually disappointed to find that their regressions yielded incorrect signs on interest rate variables approximately half the time, although these results often went unreported. Studies based on more recent data, of which that by Irvine (1981) may be the most notable, have found stronger indications that monetary conditions influence desired inventory stocks. If the great inflation of the 1970s sensitized business enterprises to the significance of interest rates as a component of inventory carrying costs, the inventory cycle may now be more of a monetary phenomenon than in the past.

See Also

- ▶ Acceleration Principle
- ▶ Business Cycles
- ▶ Cobweb Theorem
- ▶ Trade Cycle

Bibliography

Abramovitz, M. 1950. *Inventories and business cycles*. New York: National Bureau of Economic Research.
Blinder, A.S., and S. Fisher. 1981. Inventories, rational expectations, and the business cycle. *Journal of Monetary Economics* 8(3): 277–304.
Chikàn, A. 1984. Inventory fluctuation in the Hungarian economy. In *New results in inventory research*, ed. A. Chikàn. New York: Elsevier.
Hawtrey, R.G. 1928. *Trade and credit*. London: Longmans, Green & Co.
Hirsch, A.G., and M.C. Lovell. 1969. *Sales anticipations and inventory behavior*. New York: Wiley.
Holt, C.C., and F. Modigliani. 1961. Firm cost structure and the dynamic responses of inventories, production, work force and orders to sales fluctuations. In *Inventory fluctuations and economic stabilization*. Washington, DC: Joint Economic Committee, US Congress.
Irvine Jr., F.O. 1981. Retail inventory investment and the cost of capital. *American Economic Review* 71(4): 633–648.
Lovell, M.C. 1962. Buffer stocks, sales expectations and stability: A multisector analysis of the inventory cycle. *Econometrica* 30(April): 267–296.
Lovell, M.C. 1974. Monetary policy and the inventory cycle. In *Trade stability and macroeconomics: Essays in honor of Lloyd A. Metzler*, ed. G. Horwich and P.A. Samuelson. New York: Academic.
Lundberg, E. 1937. *Studies in the theory of economic expansion*. London: P.S. King & Sons.
Metzler, L. 1941. The nature and stability of inventory cycles. *Review of Economic Statistics* 23: 113–129.
Pigou, A.C. 1929. *The function of economic analysis*. London: Oxford University Press.

Inventory Investment

James A. Kahn

Abstract

Interest in inventory investment's role in business cycle volatility goes back at least to John Maynard Keynes. This article examines some basic facts about aggregate inventory investment, emphasizing its highly volatile and pro-cyclical nature. It then outlines several approaches to modelling inventory behaviour, including a detailed discussion of the linear-quadratic model, and examines their implications for inventory investment's potential role in business cycle fluctuations. The article concludes with a discussion of the potential for progress in inventory control methods to have played a role in the decline in aggregate volatility since the mid- 1980s.

Keywords

Adjustment costs; Business cycles; Credit constraints; Dynamic programming; Flexible accelerator models; Inventory behaviour; Inventory investment; Linear-quadratic models; Metzler, L. A.; National income accounts; Non-convexity; s-S models; Stockout-avoidance model

JEL Classifications

D4; D10

Inventory investment is the change in the stocks of materials, works in process, and finished goods within a firm, industry, or entire economy over a specified period of time. Because in most instances the measure encompasses a variety of goods, it is usually measured in currency units, perhaps deflated (for example, in 1999 dollars). Occasionally, however, when highly disaggregated data are available, it can be measured in physical units (for example, Blanchard 1983; Kahn 1992).

In national income accounts, aggregate inventory investment is the difference between Gross Domestic Product (GDP) and final sales of domestic product. As a share of GDP it is tiny but highly volatile in modern industrial economies. In the post-war United States, for example, it averages 0.62 per cent of GDP, but has a standard deviation of 0.83 per cent. By comparison, fixed non-residential investment averages 10.6 per cent of GDP with a standard deviation of 1.2 per cent. (Data for these calculations come from the US National Income and Product Accounts, Table 5.)

Inventory investment is also highly pro-cyclical. For example, its correlation with real GDP growth in post-war US data is approximately

0.4, and very close to the correlation between fixed non-residential investment and real GDP growth. Also, the standard deviation of real GDP growth is substantially higher than that of final sales (4.0 versus 3.3 per cent), notwithstanding the fact that for more than half of the economy GDP and final sales are identical. Thus inventory investment 'adds' to the volatility of GDP growth in the accounting (though not necessarily causal) sense. Indeed, interest in inventory behaviour as a contributor to aggregate volatility goes back at least to Keynes (1936), and includes notable contributions by Metzler (1941) and Abramovitz (1950). Blinder (1981, p. 500) writes that 'to a great extent, business cycles are inventory fluctuations'.

The pro-cyclicality of inventory investment appears inconsistent with standard microeconomic models of inventory behaviour, particularly those that stress 'buffer stock' or 'production-smoothing' motives, as noted by, among others, Blinder (1986) and West (1986). And this was not the first puzzle brought to light by research on inventory behaviour. Some ten years earlier, Feldstein and Auerbach (1976) noted the persistence of inventory-sales ratios' deviations around their means (see also Ramey and West 1999), particularly given the trivial adjustments needed to restore them to a (presumed) fixed target.

Researchers have also found inventory behaviour informative about the fundamental driving forces of business cycles (see West 1990). For example, Blinder (1986), Eichenbaum (1989), Kydland and Prescott (1982) and Christiano (1988) hypothesize supply side disturbances to account for pro-cyclical inventory investment. Others (such as Ramey 1991; Hornstein and Fisher 2000) consider non-convexities such as fixed costs or downward-sloping marginal cost. Kashyap et al. (1994) argue for the importance of credit constraints. By contrast, Bils and Kahn (2000) argue that the counter-cyclical behaviour of inventory-sales ratios casts doubt on such supply side explanations, which imply counterfactually that inventories should be relatively tight (in relation to sales) during recessions and plentiful in expansions.

The Linear-Quadratic Model

The workhorse of applied inventory research is the linear-quadratic cost minimization model developed by Holt et al. (1960). The firm is assumed to face a stochastic demand process independent of its inventory and production decisions. Consequently, whether it is a competitive price-taker or has monopoly power, the firm can condition on its expected sales process and minimize costs, which take the form

$$E_t\left\{\sum_{\tau=t}^{\infty}\beta^{\tau-t}\left[c_1 y_\tau + c_2 y_\tau^2 + c_3\left(h_\tau - h_\tau^*\right)^2\right]\right\}, \tag{1}$$

subject to

$$h_\tau = h_{\tau-1} + y_\tau - s_\tau, \tag{2}$$

where y denotes production, s sales, h the end-of-period inventory stock, h^* the desired or 'target' stock, and β a discount factor. Some versions of the model include additional cost terms such as a cost of changing production. The target h^* is usually assumed to be either a constant or proportional to expected sales. In addition, c_1 may be stochastic, and there may be additional additive stochastic terms (for example, materials prices).

A standard informational assumption is that production decisions at date t are based on period $t-1$ information, with the implication that h_t is not controlled directly. Letting $\theta = c_2/c_3$, the solution to the problem (1) and (2) is:

$$E_{t-1}\{h_t\} = \lambda h_{t-1} + \lambda E_t\left\{\sum_{\tau=t}^{\infty}(\beta\lambda)^{\tau-t}\left[\theta^{-1}h_\tau^* + \beta s_{\tau+1} - s_\tau\right]\right\} \tag{3}$$

where $\lambda \in (0, 1)$; 1Þis the smaller root of $\beta\lambda^2 - (1 + \beta + \theta^{-1})\lambda + 1$. In the limiting case with $\theta = c_2 = 0$ the solution is $E_{t-1}\{h_t\} = h_t^*$. If h^* is a constant, then the only motive for varying inventories is to smooth production. Durlauf and Maccini (1995) decisively reject this version of the model. It is worth noting that the solution (3) bears

some similarity to another widely used model, the flexible accelerator of Lovell (1961) and others

$$h_t - h_{t-1} = \lambda_1\left(h_t^* - h_{t-1}\right) - \lambda_2(s_t - E_{t-1}\{s_t\}) + u_t,$$

where u_t is a disturbance term, and both λ_1 and λ_2 lie between 0 and 1.

Both the linear-quadratic and flexible accelerator models, however, have a history of empirical difficulties. Blinder (1986) pointed out that, for the pure production-smoothing model (with h^* a constant), the model counterfactually implies that the variance of sales exceeds that of production. West (1986) showed that a more general variance inequality implied by the model with a target proportional to expected sales is also violated in US manufacturing data. Moreover, among studies of similar data, there is disagreement in the literature on the magnitudes, and even the signs, of key parameters. For example, Ramey (1991) finds negatively sloped marginal cost, in contrast to most other studies. West (1986) finds a relatively small cost of inventory deviations from their target. West and Wilcox (1994) find that obtaining precise estimates of the linear-quadratic model may be problematic with realistic sample sizes.

Regarding the flexible accelerator, Feldstein and Auerbach (1976) estimated small values for both λ_1 and λ_2, which is paradoxical because a small λ_1 implies large adjustment costs, but a small λ_2 implies that sales surprises are largely offset by within-period production responses. Their proposed solution is a target ratio that itself adjusts slowly over time. They do not provide a strong theoretical foundation for their 'target adjustment' model, however. One theme of the alternative approaches discussed in the next section is the effort to base inventory models on more rigorous microfoundations in the hope of resolving the empirical puzzles.

Other Approaches

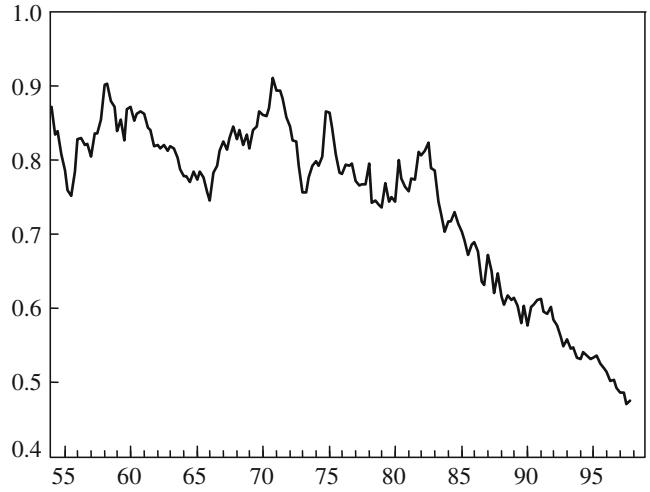
Motivated by the empirical difficulties described above, researchers have examined a number of alternative approaches to modelling inventory behaviour. One, the socalled 'stockout-avoidance' model, provides a rigorous microfoundation for the target stock. Building on Karlin and Carr (1962), Kahn (1987, 1992) considers a firm that faces a non-negativity constraint on its inventories, and must commit to production and pricing decisions each period before observing potential sales, or 'demand' x_t. Consequently, sales equal the minimum of x_t and the stock available $h_{t-1} + y_t$. If we let F denote the distribution function for x, profit maximization implies

$$p_t(1 - F(h_{t-1} + y_t)) - c_t + \beta E_t\{c_{t+1}\}F(h_{t-1} + y_t) = 0,$$

where p is price and c is marginal cost. Then, if demand uncertainty is multiplicative, for example, and p and c (and hence the markup) are constant, the firm will set $h_{t-1} + y_t$ proportional to expected demand $E_{t-1}\{x_t\}$. In addition, positive serial correlation in demand results in the variance of production exceeding the variance of sales.

Another important implication of this approach is that inventory-sales ratios depend on price–cost markups. Bils and Kahn (2000) show, in a model in which expected sales are increasing in the stock available, that the optimal inventory–sales ratio is a function of the markup and a discount rate $\beta E_t\{c_{t+1}/c_t\}$. They argue that the counter-cyclical behaviour of the inventory–sales ratio implies a counter-cyclical markup, or, equivalently, pro-cyclical marginal cost.

An alternative approach builds on the work of Scarf (1960), who modelled inventory behaviour with fixed ordering costs. Scarf provided conditions under which inventories would fluctuate between a fixed upper and lower bound, which he dubbed 'S' and 's' respectively – hence the moniker (S, s) model. (The conditions, such as i.i.d. orders, are quite restrictive, however.) Caplin (1985) showed that this model implies that the variance of orders exceeds the variance of sales, and Hornstein and Fisher (2000) extended this approach to a general equilibrium setting. Hall and Rust (2000) provide some empirical support for 'generalized' (S, s) behaviour



Inventory Investment, Fig. 1 Inventory-sales ratio, durable goods, USA, 1954–1998. Note: Inventories and sales are in chained 2000 dollars (*Source*: US National Income and Product Accounts. Durable goods inventories are from Table 5.7.6A, and final sales are from Table 1.2.6)

where the two limits depend on the spot price of the good in inventory. While the existence of fixed costs at the microeconomic level is well established, their importance for aggregate inventory behaviour at business cycle frequencies remains a matter of debate.

Inventories and the Great Moderation

Recently attention has again turned to inventory behaviour as a possible explanation for the dramatic reduction in aggregate volatility, which in the United States dates from approximately 1984 (McConnell and Perez-Quiros 2000). Kahn et al. (2002) show that reduced volatility is most pronounced in the durable goods sector, and for production more than for sales. At the same time, that sector has experienced large declines in inventory–sales ratios, as shown in the accompanying Fig. 1, and reduced volatility of inventory investment. They also provide a model in which improved information about demand shocks results in reduced output volatility. While there is much anecdotal evidence of efforts to improve inventory control by techniques such as 'just-in-time' management, there remains nonetheless considerable debate over the importance of inventories in increased aggregate stability.

See Also

- ▶ Dynamic Programming
- ▶ s-S Models

Bibliography

Abramovitz, M. 1950. *Inventories and business cycles*. New York: National Bureau of Economic Research.

Bils, M.J., and J.A. Kahn. 2000. What inventory behavior tells us about business cycles. *American Economic Review* 90: 458–481.

Blanchard, O.J. 1983. The production and inventory behavior of the U.S. automobile industry. *Journal of Political Economy* 91: 365–400.

Blinder, A.S. 1981. Retail inventory behavior and business fluctuations. *Brookings Papers on Economic Activity* 1981(2): 443–505.

Blinder, A.S. 1986. Can the production smoothing model of inventory behavior be saved? *Quarterly Journal of Economics* 101: 431–453.

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

Christiano, L. 1988. Why does inventory investment fluctuate so much? *Journal of Monetary Economics* 21: 247–280.

Durlauf, S., and L. Maccini. 1995. Measuring noise in inventory models. *Journal of Monetary Economics* 36: 65–90.

Eichenbaum, M. 1989. Some empirical evidence on the production level and production cost smoothing properties of inventory investment. *American Economic Review* 79: 853–864.

Feldstein, M.S., and A. Auerbach. 1976. Inventory behavior in durable goods manufacturing: The target adjustment model. *Brookings Papers on Economic Activity* 1976(2): 351–396.
Hall, G., and J. Rust. 2000. An empirical model of inventory investment by durable commodity intermediaries. *Carnegie-Rochester Conference Series on Public Policy* 52: 171–214.
Holt, C.C., F. Modigliani, J.F. Muth, and H. Simon. 1960. *Planning production, inventories, and work force*. Englewood Cliffs: Prentice-Hall.
Hornstein, A., and J. Fisher. 2000. (S, s) inventory policies in general equilibrium. *Review of Economic Studies* 67: 117–145.
Kahn, J.A. 1987. Inventories and the volatility of production. *American Economic Review* 77: 667–669.
Kahn, J.A. 1992. Why is production more volatile than sales? Theory and evidence on the stockout-avoidance motive for inventory-holding. *Quarterly Journal of Economics* 107: 481–510.
Kahn, J.A., M. McConnell, and G. Perez-Quiros. 2002. On the causes of increased stability in the US economy. *Economic Policy Review* 8: 183–202.
Karlin, S., and C.R. Carr. 1962. Prices and optimal inventory policy. In *Studies in Applied Probability and Management Science*, ed. K. Arrow, S. Karlin, and H. Scarf. Stanford: Stanford University Press.
Kashyap, A., J. Stein, and D. Wilcox. 1994. Credit conditions and the cyclical behavior of inventories. *Quarterly Journal of Economics* 109: 565–592.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. London/New York: Macmillan/Harcourt Brace.
Kydland, F., and E. Prescott. 1982. Time-to-build and aggregate fluctuations. *Econometrica* 50: 1345–1370.
Lovell, M.C. 1961. Manufacturers' inventories, sales expectations, and the acceleration principle. *Econometrica* 29: 293–314.
McConnell, M., and G. Perez-Quiros. 2000. Output fluctuations in the United States: What has changed since the early 1980s? *American Economic Review* 90: 1464–1576.
Metzler, L. 1941. The nature and stability of inventory cycles. *Review of Economics and Statistics* 23: 113–129.
Ramey, V. 1991. Nonconvex costs and the behavior of inventories. *Journal of Political Economy* 99: 306–334.
Ramey, V., and K. West. 1999. Inventories. In *The handbook of macroeconomics*, ed. J.B. Taylor and M. Woodford. Amsterdam: Elsevier Science.
Scarf, H. 1960. The optimality of (S, s) policies in the dynamic inventory problem. In *Mathematical methods in the social sciences*, ed. K. Arrow, S. Karlin, and H. Scarf. Stanford: Stanford University Press.
US National Income and Product Accounts. 2007. Online. Available at http://www.bea.gov/bea/dn/nipaweb/SelectTable.asp. Accessed 23 Feb 2007.
West, K.D. 1986. A variance bounds test of the linear quadratic inventory model. *Journal of Political Economy* 94: 374–401.
West, K.D. 1990. The sources of fluctuations in aggregate inventories and GNP. *Quarterly Journal of Economics* 105: 939–972.
West, K.D., and D. Wilcox. 1994. Estimation and inference in the linear-quadratic inventory model. *Journal of Economic Dynamics and Control* 18: 897–908.

Inventory Policy Under Certainty

Arthur F. Veinott Jr.

Inventories of raw materials, work-in-process, and finished goods are ubiquitous in firms engaged in production and/or distribution of one or more products. Indeed, in the United States alone, 1982 non-farm business inventories totalled over 500 billion dollars, or about 17 per cent of the gross national product that year. The annual cost of carrying these inventories, e.g., costs associated with capital, storage, taxes, insurance, etc., is significant – perhaps 25 per cent of the total investment in inventories, or about 125 billion dollars. Since the cost of carrying inventories is sizeable, a good deal of attention has been devoted to the problem of determining optimal or near-optimal inventory policies that properly balance the costs and benefits of carrying inventories. Moreover, since a firm's inventories are usually distributed among several facilities, e.g., plants, warehouses, retail outlets, effectively coordinating the inventory policies in multi-facility systems.

This essay has four goals. One is to discuss some of the main motives for carrying and distributing inventories in multifacility inventory systems under certainty. Another is to explain how the form of optimal and/or near-optimal multi-facility inventory policies depend on the particular motive(s) present for carrying and distributing inventories. The third is to outline how the special structure of multi-facility inventory models can be exploited to carry out computations efficiently. The last is to give a brief historical perspective on some of the main

developments in multi-facility inventory policy under certainty.

Since it is usually expensive to carry inventories, efficient firms would not do so without good reasons. Thus, it seems useful to ask what motivates a firm to carry inventories at a facility. For this purpose, it is convenient to differentiate *retailers*, i.e., facilities that face exogenous demands, from *wholesalers*, i.e., facilities that face only endogenous demands. Moreover, we shall assume that each facility produces a single product and that each unit output thereof consumes fixed amounts of the outputs of the other facilities that ship to it. Production at a facility is intended to have a broad interpretation, including procurement and shipments from another facility. There are known multi-period non-negative exogenous demands for the output of each facility in excess of the endogenous demands generated by the other facilities. There are temporally varying production and storage costs at each facility that depend on the respective amounts produced and stored there. The goal is to find production and storage schedules at each facility that minimize the total system cost.

Motives for Retailers to Carry Inventories

Under these circumstances, two of the most significant motives for a retailer to carry inventories are the following.

Scale Economies in Supply

Scale economies in supply occur for several reasons including procurement quantity discounts or set-up costs and production/transportation scale economies. When that is so, it is often economical to produce in a single period to satisfy exogenous and endogenous demands occurring over several periods with the aim of reducing the average unit cost of supply.

Temporal Increase in Marginal Cost of Supplying Demand

The marginal cost of supplying demand in a period is the marginal cost of producing an amount equalling the demand in the period. A temporal increase in the marginal cost of supplying demand may arise in several ways. One is where the demands are stationary and there is a temporal increase in the marginal cost of production. The latter occurs, for example, when raw-materials prices, wage rates, or marginal transportation costs increase. A second is where there is a temporal increase in demand and diseconomies of scale in production. A temporal increase in demand may arise because of long-term growth or seasonality thereof. Production scale diseconomies occur when there are alternate sources of supply, each with limited capacity, or when production at a plant in excess of normal capacity must be deferred to a second shift or to over-time with an attendant increase in unit labour costs.

Motives for Wholesalers to Carry Inventories

Each of the above motives for retailers to carry inventories must be strengthened in order to motivate wholesalers to do likewise. Three significant ways to do this for a wholesaler whose output is directly or indirectly consumed by some retailer are the following.

Inter-Facility Storage-Cost Variations

Inter-facility variations in storage costs are common, e.g., between retailers located on expensive city land and wholesalers located on inexpensive rural land. Such variations, when coupled with one of the motives for retailers to carry inventories, may also motivate wholesalers to carry inventories. To see why, consider the extreme case of a retailer with high enough storage costs to make storage there uneconomical. In that event, the retailer passes its demands on directly to any wholesaler whose output the retailer consumes – effectively making the wholesaler a retailer. Then, if either of the two motives for retailers to carry inventories is present at the wholesaler, the latter may be motivated to carry inventories. Of course, in the case of the second motive, the demand at the wholesaler is the demand passed on to it by the retailer.

Inter-Facility Variations in Supply Scale Economies

Facility-dependent supply scale economies may motivate wholesalers to carry inventories. For example, if a wholesaler and a retailer are suppliers of a second retailer and there are no production costs at either retailer, then scale economies in production at the wholesaler may motivate it to carry inventories for the second retailer.

Inter-Facility Variations and Temporal Increase in Marginal Cost of Supplying Demand

Inter-facility variations in the marginal cost of supplying demand arise because of inter-facility variations in demands or production/transportation costs, capacity limitations, etc. Such variations, when coupled with a temporal increase in the marginal cost of supplying demand, may motivate a wholesaler to carry inventories for a retailer that consumes its output. This can occur when the retailer's marginal production costs are temporally non-increasing and either there is a temporal increase in the marginal production cost at the wholesaler or there are diseconomies of scale in production at the wholesaler and rising demands at the retailer.

Formulation of Multi-Facility Inventory Problem

In order to see how the different motives for carrying inventories influence the form of desirable inventory policies, it is necessary to formulate the problem more precisely. To that end, consider a collection of *facilities*, labelled 1,. . ., *f*, each producing a single product. Facilities can be interpreted in many ways, e.g., as plants, warehouses, retail outlets, machines in a plant, etc. the facilities are linked by the fact that the production of each unit at facility *j directly consumes* $e^{ij} \geq 0$ units of the output of facility *i*. The time-lags in shipments between facilities are negligible, i.e., production at facility *j* in a period consumes output at other facilities that may be produced at those facilities either before *or during* the period. There is a given exogenous non-negative demand d_{jt} in period $t = 1,\ldots, p$ for the ouput of facility *j*. The demands at each facility in each period are met as they occur. There is a real-valued cost $c_{jt}(w)$ (resp., $h_{jt}(w)$ of producing (resp., storing) $w \geq 0$ in (resp., at the end of) period *t* at facility *j*. The cost of producing *w* units at facility *j* in a period also includes the costs of transporting e^{ij} w units of the output of each facility *i* to facility *j* in that period. We can and do assume without loss of generality that all $c_{jt}(0) = h_{jt}(0) = 0$. Let x_{jt} and y_{jt} be the respective amounts produced in and stored at the end of period *t* at facility *j*. Let $x_t = (x_{jt})$, $y_t = \left(y_{jt}\right)$ and $d_t(d_{jt})$ be respectively the *f*-element column vectors of *production. inventory and demand schedules in period t*, and let $x = (x_t)$, $y = (y_t)$ and $d = (d_t)$ be the $f \times p$ matrices of *p-period production, inventory and demand schedules*. Assume that $y_0 \equiv y_p \equiv 0$. The net production at each facility that is available to satisfy exogenous demands in, or to store at the end of, period *t* is $(I - E)x_t$ where $E \equiv (e^{ij})$ is the *consumption matrix*. The problem is to find a *production and storage schedule* $z = (x, y) \geq 0$ that minimizes the *cost*

$$\sum_{j,t} \left[c_{jt}\left(x_{jt}\right) + h_{jt}\left(y_{jt}\right)\right] \tag{1}$$

subject to the *stock-conservation constraints*

$$(I - E)x_t + y_{t-1} - y_t = d_t, \quad t = 1, \ldots, p. \tag{2}$$

Facility Network

Associated with the consumption matrix *E* is a *facility network* **F** whose nodes are the facilities and whose arcs are the ordered pairs $i \to j$ of facilities *i, j* for which $e^{ij} > 0$. There is no loss of generality in assuming that **F** is *connected*, i.e., there is an undirected path from each facility to each other facility. For if not, the problem can be solved separately for the set of facilities in each connected component, i.e., maximal connected subnetwork, of **F**.

Facility *j directly* (resp., *indirectly*) *consumes* the output of facility *i* if there is a chain, i.e., directed path, from *i* to *j* with exactly one arc

(resp., two or more arcs). Assume throughout that the facility network is *circuitless*, i.e., no facility directly or indirectly consumes its own output.

Several practical examples of circuitless facility networks are also *trees*, i.e., there is a unique simple path between each pair of facilities. Among the common special cases of trees are assembly, distribution, assembly-distribution, star and series networks. *Assembly* (resp., *distribution*) networks are rooted trees in which all arcs are directed towards (resp., away from) a distinguished facility called the root. The root is the facility at which final assembly of all products takes place in assembly networks and is the ultimate source of all product in distribution networks. *Star* networks are rooted trees in which all arcs are incident to the root, and so are assembly-distribution networks. *Series* networks are chains and so are at once assembly and distribution networks. Finally, *assembly-distribution* networks are rooted trees in which the root facility divides the tree into two subtrees, one an assembly network and the other a distribution network, both sharing the root facility. In trees, we can and do assume without loss of generality and without further mention that all products are measured in common units, i.e., $e^{ij} = 1$ for all arcs $i \to j$.

Linear Costs

In this section, we consider the multi-facility inventory problem in its simplest setting, namely, where there are neither economies nor diseconomies of scale, so the cost function (1) is linear. This allows temporal increases in unit production costs and interfacility variations in unit storage or production costs. The former may motivate retailers to carry inventories and, when coupled with one of the latter, may motivate wholesalers to do likewise.

Extreme Schedules of Totally-Leontief-Substitution Systems

The problem is a totally-Leontief-substitution system. Moreover, at least one optimal schedule is an extreme point of the set of feasible schedules, and each extreme schedule satisfies

$$y_{j,t-1}x_{jt} = 0 \tag{3}$$

for $1 \le j \le f$ and $1 \le t \le p$, i.e., facility j produces in period t only if that facility has no entering inventory in the period. This is so because production at a facility in a period and storage there in the previous period are, in the terminology of Leontief-substitution systems, 'substitute' activities, and extreme schedules in such systems do not admit substitute activities. The condition (3) assures that whenever a facility produces, it satisfies all the endogenous and exogenous demands for its output in an interval of periods. Thus, there is no 'lot splitting', i.e., the endogenous demand at one facility created by production at another facility in a period is entirely satisfied by production at the first facility in a *single* (no later) period. In this sense, each facility produces in 'larger' lots than do its followers.

Dynamic-Programming Computation of Optimal Dual Prices

Moreover, the optimal dual price C_{jt} associated with the stock-conservation constraint for facility j in period t is the minimum cost of satisfying a unit of demand at facility j in period t. It can be shown that the C_{jt} satisfy, and can be calculated from, the dynamic-programming recursion

$$C_{jt} = \min\left(h_{j,t-1} + C_{j,t-1}, c_{jt} + \sum_{i \to j} e^{ij} C_{it}\right) \tag{4}$$

for $1 \le j \le f$ and $1 \le t \le p$ where c_{jt} and h_{jt} are respectively the unit product and storage costs at facility j in period t and $C_{j0} \equiv h_{j0} \equiv 0$. Equation expresses the fact that the optimal way to satisfy a unit of demand at facility j in period t is to choose the cheaper of two options. One is to provide a unit of product at facility j in period $t-1$ as cheaply as possible and store it for one period. The other is to produce a unit at facility j in period t, thereby consuming e^{ij} units from each facility i in that period with those units being provided as cheaply as possible. Observe that it is optimal to produce a facility j in period t if and

only $C_{jt} \leq h_{j,t-1} + C_{j,t-1}$,independently of the demand schedule d.

Once the optimal periods in which to produce each product are found, the desired optimal production schedule is obtained inductively as follows. Suppose the optimal production schedules $\mathbf{x}^j = (\mathbf{x}_{jt})$ have been found for all facilities j that directly consume the output of facility i. Then, if it is optimal for facility i to produce in period t and next in period $u + 1$, it follows from (2) and (3) that x_{it} can be calculated recursively from

$$\mathbf{x}_{it} = \sum_{k=t}^{u} d_{ik} + \sum_{i \to j} e^{ij} \sum_{k=t}^{u} \mathbf{x}_{jk}. \qquad (5)$$

Form of Optimal Production Schedule

To sum up, some optimal schedule is extreme, and so a facility produces in a period only if the facility has no entering inventory in the period. If the consumption matrix and demand schedule are integer, then so is each extreme schedule. Optimal production at a facility in a period is a nondecreasing linear function of the present and future demand at the facility and all facilities that consume its output.

Running Time

The running time of this dynamic-programming algorithm is $O(f^2p)$ in general networks and falls to $O(fp)$ in planar networks, the last because the number of arcs in such networks does not exceed three times the number of nodes. Of course trees, and in particular. assembly-distribution networks, are planar, so the improved running time applies in such networks.

Nonlinear Costs

The algorithm given above for solving the problem with linear costs is also useful when the costs are nonlinear. This is because many methods for solving problems with nonlinear costs, e.g., branch-and-bound, gradient methods, etc., entail solving a sequence of linear-cost problems, each of which can be solved in linear time by the recursion (4).

Concave Costs

In this section we generalize the linear-cost multi-facility inventory problem discussed in Section Linear Cost to allow economies-of-scale by requiring the cost function (1) to be concave. Then the marginal costs of production and storage fall respectively the more one produces and stores. Since the class of additive concave cost functions contains the linear ones, all motives for carrying inventories at retailers and wholesalers with linear cost functions remain in force in this section. Beyond these, the introduction of scale economies in this section provides an added motive for carrying inventories at retailers. This fact, when coupled with either inter-facility variations in storage or production costs, both of which are allowed in this section, provide additional motives for carrying inventories at wholesalers.

Extreme Schedules

As for the case of linear costs, if the minimum is attained, it is attained at an extreme schedule and so satisfies (3). However, because of the scale economies, optimal production is no longer linear in demand. For that reason, the dynamic-programming recursion (4) no longer solves the problem.

Dynamic-Programming Algorithm for Series Networks

However, there is a polynomial-time dynamic-programming algorithm for finding an optimal schedule in series networks in which facility f is the only retailer. To describe the algorithm, observe first that on iterating a representation like (5) of any extreme schedule $z = (x,y)$, that the sum $y_{i,t-1} + x_{it}$ of the initial inventory and production at facility i in period t equals the sum $d^{jl} \equiv \sum_{k=j}^{l-1} d_{jk}$ of the demands at facility f in periods $j,\ldots, l-1$ for some $t \leq j < l$. Let $C_{it}(d^{jl})$ be the minimum cost of satisfying the demands at facility f in periods $j,\ldots, l-1$ from the stock d^{jl} available at facility i at the beginning of period t. Then the $C_{it}(d^{jl})$ can be calculated from the dynamic-programming recursion.

$$C_{it}(d^{jl}) = \min_{\substack{j \le k \le l \\ t<k}} l\left[c_{i+1,t}(d^{jk}) + C_{i+1,t}(d^{jk}) + h_{it}(d^{kl}) + C_{i,t+1}(d^{kl})\right]$$

for $1 \le i < f$, $1 \le t \le j < l \le p + 1$ and $t < p$, together with fairly obvious boundary recursions where $i = f$ or $t = p$. This recursion expresses the fact that the minimum cost of satisfying the demands at facility f in periods $j,\ldots, l-1$ from the stock d^{jl} on hand after production at facility i in period t is attained by dividing the stock d^{jl} into two parts, d^{jk} and d^{kl}, the former being sent to facility $i + 1$ for production in period t and the latter being stored at facility i for one period.

Running Time

The running time of the algorithm is $O(fp^4)$, which is $O(p^3)$ times that for the linear-cost case. If also the production and storage costs at each facility are respectively temporally non-increasing and non-decreasing in the facility are respectively index, the running time improves to $O(fp^3)$ because some optimal schedule is 'nested' in the sense to be defined shortly. For the case of a single facility, the running time drops further to $O(fp^2)$.

General Networks

The above algorithm can be generalized to arbitrary distribution networks, but the computational effort grows exponentially with the number of retailers. For that reason, we do not discuss this possibility. For the special case of one-warehouse multi-retailer networks (i.e., star distribution networks in which the root facility is a warehouse and the other facilities are retailers) in which the production and storage costs at the warehouse are respectively linear plus a set-up cost and linear, there is an algorithm for solving the problem whose running time is linear in the number of retailers, but exponential in the number of periods. However, no polynomial-time algorithm has been found for the general multi-facility problem, even for star networks.

Effective Heuristics

This suggests the possibility that optimal schedules may be too difficult to find and that heuristics may instead be necessary. One heuristic for distribution systems is to optimize over the subclass of *nested* schedules, i.e., schedules x satisfying (3) for which $x_{it} > 0$ implies $x_{jt} > 0$ for each facility j that directly consumes the output of facility i. There is a dynamic-programming algorithm for finding an optimal nested schedule in a distribution system in $O(fp^3)$ time. Unfortunately, optimal schedules need not be nested – even in one-warehouse multi-retailer networks – because, for example, it may be optimal for a retailer with low demands to order less frequently than the warehouse. However, the nested-schedules heuristic can be adapted to give a reasonably effective heuristic when the demand schedules at each facility are *proportional* to one another.

Stationary Case with Set-Up Production and Linear Storage Costs

It turns out that there is an extraordinarily effective heuristic for the stationary (demands and costs) continuous-time infinite-horizon version of the problem in which there is a set-up production cost and a linear storage cost at each facility. The *effectiveness* of a heuristic for this problem is 100 per cent times the ratio of the infimum of the average cost per unit time over all policies to the average cost per unit time incurred by the heuristic.

The heuristic for this problem decomposes the multi-facility problem into a collection of single-facility problems, one for each facility. In each single-facility problem, there is a demand rate $r > 0$ per unit time, a set-up production cost $K > 0$ and a storage cost $h > 0$ per unit stored per unit time. Then one expects that a minimum-average-cost schedule will permit production only when stock runs out, which is the continuous-time analogue of (3). Thus, since the demand rates and costs are stationary, one anticipates that a minimum-average cost schedule will entail producing every $T > 0$ periods an amount equalling the demand rT until the next instance of production. An easy calculation shows that the optimal value of T is given by the celebrated square-root formula $T^o = (2K/hr)^{1/2}$. Now if $T > 0$ is any other production interval for which $2^{-1/2} \le T/T^o \le 2^{1/2}$, then it is easy to show that the effectiveness of the new schedule is at least 94 per cent.

The heuristic for the multi-facility problem is constructed as follows. First, restrict attention to *power-of-two* schedules, i.e., those for which a facility produces only when it runs out of stock, the *production intervals* between successive times that a facility produces are identical, and the ration of the production intervals at distinct facilities is a (possibly negative) integer power of two. It is not difficult to find an expression $C(T)$ for the average cost of a power-of-two schedule $\mathbf{T} = (T^i)$ where T^i is the production interval used at facility i.

The *optimal power-of-two problem* is that of finding $\mathbf{T}$ that minimizes $C(\mathbf{T})$ subject to the power-of-two constraints and $\mathbf{T} \gg 0$. Instead of solving this problem, one next solves the *relaxation* thereof in which the power-of-two constraints are dropped. The relaxation is a minimum-convex-cost dual network-flow problem on a *cost network*. The nodes of the cost network correspond to the chains in the facility network that end at retailers. The arcs in the cost network join each node α to the node β that corresponds to the subchain formed by deleting the first node of the chain that corresponds to α. Denote by $\mathbf{T}^* = (T^{*i})$ the optimal solution of the relaxation of the optimal power-of-two problem. Remarkably and most important, the minimum average cost $C(T^*)$ of the relaxation is a lower bound on the average cost of an *arbitrary* (not necessarily power-of-two) schedule!

The cost-network problem entails reallocating the set-up costs and storage cost rates among the facilities in such a way that with the new cost parameters, each facility i's optimal production interval, when considered as a single-facility problem, is precisely T^{*i}. One then rounds off each T^{*i} to form a power-of-two schedule $\mathbf{T} = (T^i)$ satisfying $2^{-1/2} \leq T^i/T^{*i} \leq 2^{1/2}$ for each i. This assures that the effectiveness of the resulting power-of-two schedule is at least 94 per cent for each facility, and so is at least 94 per cent for the system. A somewhat more complex procedure guarantees that the system effectiveness is at least 98 per cent.

The cost-network problem can be reduced to solving a sequence of maximum-flow problems, each of which splits its predecessor into two smaller subproblems. In special cases, there are even more efficient algorithms. For example, the cost-network problem can be solved in $O(f \ln f)$ time in one-warehouse multi-retailer and assembly networks.

Convex Costs

In this section we discuss the multi-facility inventory problem in the presence of diseconomies of scale by requiring the cost function (1) to be convex. Then the marginal costs of production and storage are non-decreasing in the amounts produced and stored respectively. This allows a temporal increase in the marginal cost of supplying demand, which may motivate retailers to carry inventories.

s-additive Convex Production Cost Function

Here we suppose that the production cost function is *s-additive convex*, i.e., $c_{ji}(w) = S_t c^j(w/s_t)$ for $w \in R$ with $c^j(\cdot)$ being a $+\infty$ or real-valued convex function on the real line and s_t being a positive *scale parameter* for each t. Also assume that there are no direct storage costs, so $h_{jt} \equiv 0$, but that it is possible to represent any such costs by absorbing them in the production costs with an appropriate choice of the scale parameters s_t, e.g., as we show below for capital costs. In particular, there is a motive to carry inventories at retailer j in period t if the marginal cost $\dot{c}(d_{jt}/st)$ of supplying the demand there in that period is less than that in some subsequent period k, say. This implies, and provided $\dot{c}^j(w)$ is strictly increasing in w, is implied by $d_{jt}/S_t < d_{jk}/S_k$. The last is so if $Od_{jt} \leq d_{jk}$ and $s_t \geq s_k$ with at least one inequality being strict. This formulation is rich enough to be useful because it provides for certain storage costs and allows temporal variations in the marginal cost of supplying demand, and yet is simple enough to admit a graphical solution.

Positively-Homogeneous Convex Production Cost Function

As a particular example, suppose that the present value of the production cost at facility j is

positively homogeneous of degree $q + 1 \geq 1$, i.e., $c_{jt}(w) = \beta^t |w|^{q+1} c_j$ for some discount factor $\beta > 0$ and $c_j > 0$. This accounts for the cost of capital invested in inventories. Then the production cost is *s*-additive convex with scale parameter

$$s_t = (1 + \rho)^{t/q} \tag{6}$$

where $\beta \equiv 1/(1 + \rho)$ and 100ρ per cent is the interest rate. Observe that if $\rho = 0$, then $s_t = 1$ for all t. If instead $\rho > 0$ (resp., $\rho < 0$), then s_t expands (resp., contracts) geometrically with the precise rate being greater than, equal to or less than $|\rho|$ according as $0 < q < 1, q = 1$ or $1 < q$.

Taut-String Solution of the Single-Facility Problem

Now return to the case of an arbitrary *s*-additive convex production cost function. Then the fundamental result for the single-facility problem is the Invariance Theorem which asserts that there is an optimal production schedule that is independent of the function $c \equiv c^1$, though it does depend on the demand schedule and scale parameters. Because of the Invariance Theorem, it suffices to solve the problem for any single strictly-convex function *c*. It turns out to be felicitous to put $c(w) = (w + 1)^{1/2}$ (which is strictly convex) because the problem can then be solved graphically. To see this, observe that the cost of a schedule (x,y) is then the length $\sum_t \left(x_t^2 + s_t^2\right)^{1/2}$ of the shortest polygonal path in the plane passing in order through the points (S_t, X_t) for $t = 0,\ldots,$ p where $1\ S_t \equiv \sum_1^t s_k$ and $X_t \equiv \sum_1^t x_k$, or equivalently, the length of a taut string passing in order through those points. Now consider the taut string passing through the points $(S_0, D_0) = (0, 0)$ and (S_p, D_p), and lying above the points (S_t, D_t) for $t = 1,\ldots, p - 1$ where $D_t \equiv \sum_1^t d_k$.Let (S_t, X_t^*) be the coordinates of the taut string corresponding to S_t for each t. Then (X_t^*) is the least concave majorant of (D_t). Since the feasible cumulative production schedules are non-decreasing and majorize (D_t), it follows from the Invariance Theorem and the non-negativity of the demands that $x_t^* = X_t^* - X_{t-1}^* \geq 0$ is optimal for all t and convex c.

The above *taut-string solution* has the property that during any interval of periods in which inventories are held, optimal production is proportional to the scale parameter. Thus, if the scale parameter rises (resp., falls) over the interval, then so does optimal production. In particular, if the scale parameter is given by (6), then optimal production rises or falls geometrically in the interval according as $\rho > 0$ or $\rho < 0$.

The taut-string solution has another property that plays an important role in solving the multi-facility problem, namely, it is positively homogeneous of degree one in the demand schedule. This is easily seen because, by the Invariance Theorem, there is no loss in taking the production cost function to be positively homogeneous.

Proportional Demand Schedules

In order to obtain a tractable solution to the multi-facility problem, we shall assume that the demand schedules at each facility are *proportional*, i.e., there is a *p*-period row vector *d** of nonnegative demands and a column vector δ of non-negative *demand levels* associated with the *f* facilities such that $d = \delta d^*$. Thus the demand schedules at each facility exhibit a common pattern of temporal variation. This includes the cases in which the demands are facility-dependent and stationary, or there is a single retailer. Let π be the column vector of amounts that would have to be produced at the *f* facilities to satisfy the column vector δ of exogenous demands at those facilities. Evidently, $\delta = (I - E)\pi$. Let *x** be the (row) optimal production schedule (the tautstring solution) for the *standard* single-facility problem with demand schedule *d**, and let *y** be the corresponding row vector of inventories.

Optimal Schedule for the Multi-Facility Problem

Let **x** and **y** be the $f \times p$ matrices of optimal *p*-period production and inventory schedules at each facility. By combining the above results and the Invariance Theorem, one finds that

optimal *p*period production and inventory schedules for the multifacility problem are given by the pair of rank-one matrices

$$z = (x, y) = (\pi x^*, \delta y^*).$$

Thus, the optimal production and inventory schedules at different facilities are proportional to one another. Indeed, inventories are held at a retailer only to satisfy its exogenous demands, and are proportional to its demand level. No inventories are held at a facility to satisfy endogenous demands there because the production schedules at the facilities consuming the output of the given facility already smooth out the demand schedules. Hence, *just-in-time* scheduling, i.e. holding no inventories, is optimal at all wholesalers. This is consistent with the absence of either of the motives for wholesalers to hold inventories, namely, interfacility variations in storage costs or marginal costs of supplying demand. The last is fundamentally because there are no inter-facility variations in the scale parameters of the production cost functions and the demand schedules at each facility are proportional. Hence, up to a monotone transformation, all facilities exhibit the same temporal variation in their marginal costs of supplying demands.

Running Time

The optimal schedule x^* for the standard single-facility problem can be found in $O(p)$ time. The vector π can be computed in $O(f^2)$ time in general networks and in $O(f)$ time in planar networks. Thus, $\mathbf{z}$ can be found in $O(f^2 + fp)$ time in general networks and in $O(fp)$ time in planar networks.

Historical Perspective

The study of inventory problems has influenced and been influenced by the tools available for their analysis. The seventy odd years since the first economic-lot-size model was proposed by F.W. Harris (1915) can be reasonably divided into three phases. During the period 1915–1950, attention was focused mainly on the formulation and closed-form solution of relatively simple single-facility models under certainty using the calculus. The period 1950–1965 saw the introduction of dynamic, linear and nonlinear programming methods for the characterization and computation of optimal policies in the presence of certainty and/or uncertainty, again largely for single-facility problems.

During the period since 1960, attention has gradually shifted towards the problem of coordinating the inventory policies at several interrelated facilities. It was realized that many such multi-facility inventory problems could be formulated as dynamic nonlinear network-flow or Leontief-substitution models and that the special structure of these models permitted a unified theory of inventory control to begin to emerge. A qualitative theory of mathematical programming, namely, lattice programming and substitutes, complements and ripples, was developed to give a simple general method of studying the qualitative variation of optimal inventory policies with the problem parameters. The realization that optimal policies for multi-facility systems could be extraordinarily complex to compute and/or implement, coupled with the successful use of heuristics in several areas of combinatorial optimization, led to the development of provably effective and efficient heuristics for multi-facility inventory problems. The rapid recent development of computational geometry will no doubt stimulate further progress on multi-facility systems as we begin to be able to combine the symmetries of continuous space-time models with the computational efficiency of discrete space-time models. We close by summarizing the main sources for the material in this entry. Motives for retailers to hold inventories are discussed by Arrow (1958) and Scarf (1963). The treatment of the multi-facility linear-cost problem is taken from Veinott (1969) who applied Dantzig's (1955) theory of Leontief-substitution systems. Earlier ad hoc treatments of the single-facility linearcost case are reviewed by Arrow (1958). Clark and Scarf (1960) solved the series-network problem with stochastic demands at the last facility and linear costs at all facilities except the first one.

I

The results for the single-facility concave-cost problem begin with Wagner and Whitin (1958). The characterization of the extreme schedules for the multi-facility problem is due to Zangwill (1966), though the development using Leontief-substitution systems is taken from Veinott (1969). The algorithm for series networks is due to Zangwill (1969). The nested-schedules algorithm for distribution systems comes from Veinott (1969). Love (1972) gave conditions assuring that the last algorithm finds an optimal schedule for series networks. Erickson et al. (1981) have recently encompassed an improvement of the above algorithms for series networks in a send-and-split method for finding minimum-concave-cost network flows. This result reveals that the polynomial running time of these algorithms for series networks is explained by the planarity of the corresponding network-flow problem. The results on the 94 per cent and 98 per cent effective power-of-two heuristics for the multi-facility problem with stationary demands and costs are due to Roundy (1985a, 1986). He employed an algorithm of Maxwell and Muckstadt (1985) to solve the cost-network problem. The effectiveness of a 'nested-schedules' heuristic with proportional demand schedules is discussed by Roundy (1985b).

The Invariance Theorem for the single-facility convex-cost problem is due to Modigliani and Hahn (1955) for the case of stationary costs. The generalization to non-stationary costs and the taut-string solution are due to Veinott (1971) who encompassed this example in a theory of Invariant Network Flows. The results for the multi-facility convex-cost problem with proportional demand schedules are due to recent research of the author.

See Also

► Operations Research

Bibliography

Arrow, K.J. 1958. Chapter I: Historical background. In *Studies in the mathematical theory of inventory and production*, ed. K.J. Arrow, S. Karlin, and H. Scarf. Stanford: Stanford University Press.

Clark, A.J., and H. Scarf. 1960. Optimal policies for a multiechelon inventory problem. *Management Science* 6(4): 475–490.

Dantzig, G.B. 1955. Optimal solution of a dynamic Leontief model with substitution. *Econometrica* 23(3): 295–302.

Erickson, R.E., C.L. Monma, and A.F. Veinott Jr. 1981. *Minimum-concave-cost network flows*. Stanford: Department of Operations Research, Stanford University. Revised (1986) as Send-and-split method for minimum-concave-cost network flows. To appear in *Mathematics of Operations Research*.

Harris, F.W. 1915. *Operation and costs* (The Factory Management Series), 47–52. Chicago: A.W. Shaw Co.

Love, S.F. 1972. A facilities in series inventory model with nested schedules. *Management Science* 18(5): 327–338.

Maxwell, W.E., and J.A. Muckstadt. 1985. Establishing consistent and realistic reorder intervals in production-distribution systems. *Operations Research* 33(6): 1316–1341.

Modigliani, F., and F.E. Hohn. 1955. Production planning over time and the nature of the expectation and planning horizon. *Econometrica* 23(1): 46–66.

Roundy, R.O. 1985a. 98%-effective integer-ratio lot-sizing for one-warehouse multi-retailer systems. *Management Science* 31(11): 1416–1430.

Roundy, R.O. 1985b. *Efficient, effective lot-sizing for multiproduct multi-stage production/distribution systems with correlated demands*, Technical Report, vol. 671. Ithaca: School of Operations Research and Industrial Engineering, Cornell University.

Roundy, R.O. 1986. A 98 %-effective rule lot-sizing for multi-product, multi-stage production/inventory systems. *Mathematics of Operations Research* 11(4): 699–727.

Scarf, H. 1963. Chapter 7: A survey of analytical techniques in inventory theory. In *Multi-stage inventory models and techniques*, ed. H. Scarf, D. Gilford, and M. Shelly, 185–225. Stanford: Stanford University Press.

Veinott Jr., A.F. 1969. Minimum concave-cost solution of Leontief substitution models of multi-facility inventory systems. *Operations Research* 17(2): 262–291.

Veinott Jr., A.F. 1971. Least d-majorized network flows with inventory and statistical applications. *Management Science* 17(9): 547–567.

Wagner, H.M., and T.M. Whitin. 1958. Dynamic version of the economic lot size model. *Management Science* 5(1): 89–96.

Zangwill, W.I. 1966. A deterministic multiproduct, multi-facility production and inventory model. *Operation Research* 14(3): 486–507.

Zangwill, W.I. 1969. A backlogging model and a multi-echelon model of a dynamic economic lot size production system – A network approach. *Management Science* 15(9): 506–527.

Investment (Neoclassical)

Robert M. Coen and Robert Eisner

Abstract

Investment is capital formation – the acquisition or creation of resources to be used in production. As such, it captures the production side of intertemporal consumption/ savings decisions. This entry focuses on neoclassical approaches to the study of investment. Theoretical and empirical issues are discussed.

Keywords

Acceleration principle; Adjustment costs; Capital–output ratio; CES production function; Depreciation; Distributed lags; Elasticity of substitution; Euler equations; Full employment; Human capital; Investment (Keynesian); Investment (neoclassical); Irreversible investment; Opportunity cost of capital; Public capital; Saving and investment; User cost of capital

JEL Classifications

E2

Investment is capital formation – the acquisition or creation of resources to be used in production. In capitalist economies much attention is focused on business investment in physical capital, such as buildings, equipment and inventories. But investment is also undertaken by governments (see public capital), nonprofit institutions and households, and it includes the acquisition of human and intangible capital as well as physical capital. In principle, investment should also include improvement of land or the development of natural resources, and the relevant measure of production should include non-market output as well as goods and services produced for sale.

Thus, acquisition of an automobile by government or households is as much investment as acquisition of an automobile by a business firm. The car is used in all cases for the production of transport services. Similarly, government construction of roads, bridges and airports is as much investment as business acquisition of trucks and planes. Expenditures for research and development are investment whether undertaken by business, government or nonprofit universities. And, most important, education and training, wherever undertaken, are major forms of investment in human capital.

There is a widespread mythology that investment is good and the more investment the better. But investment may be good or bad and there may be too much as well as too little.

Classical and neoclassical economists have stressed the role of investment in providing for the future. Maintaining the current level of output requires keeping up the existing means of production. Economic growth, or the increase in the rate of output, is then seen as depending considerably on the acquisition of additional means of production, that is, investment in excess of the wearing away or depreciation of existing capital. Investment may also contribute to higher output where the new capital 'embodies' new and improved technology. That investment will contribute to economic growth presupposes, however, that the additional capital is useful. It must have a positive net product, which is to say that the additional capital must contribute more to future production than the value of the resources used to create it.

How far one should go in allocating resources to investment depends upon our preferences for current consumption versus future consumption, or our preferences between our own consumption and that of our children and grandchildren. It also depends on the production function, that is, the terms under which additional capital can be converted into additional future output. It would hardly seem desirable to sacrifice 100 dollars of current consumption to produce 100 dollars of capital that would result in future production of only 90 dollars. The notion that this is not a relevant issue stems from the assumption that profit-seeking entrepreneurs would not freely undertake investment in which the costs are greater than the returns. It is not always perceived, however, that where governments offer subsidies

or 'tax incentives' for businesses to undertake investment that would not otherwise seem profitable, such unproductive capital formation is exactly what may be expected.

A second major role for investment has been seen in the achievement and maintenance of full employment. This requires that aggregate investment plus aggregate consumption equal the total output that would be produced if all individuals who wish to work could find employment. Investment may then be inadequate not only in failing to provide sufficient resources for future production, it may also be inadequate if it is insufficient to bring about the full utilization of existing resources. This latter problem has received major attention as a consequence of the work and influence of John Maynard Keynes (1936).

Another way of stating the condition necessary for full employment is that aggregate investment must equal aggregate saving out of the full-employment level of income. In national income accounts, measured investment and saving are always identically equal, owing to the identity of output and income, which, apart from receipts from abroad, is earned only from production. That part of income not spent on consumption is saved. But that part of production not purchased by consumers must be acquired (or kept) by producers and hence is investment, though not necessarily intended investment. If we designate Y as income and output, C as consumption, S as saving and I as investment, we then have $S = Y - C = I$.

While realized investment is thus identically equal to saving, investment and saving may be more or less than investment demand, that is, intended investment. If investment demand is less than saving at the current level of income, producers will find that they cannot sell all that they produce. They will accumulate undesired inventories of finished goods (unintended investment in inventories), which should lead them to reduce production. Reduced production means less income and hence less consumption and saving. A shortfall of investment demand in relation to saving therefore brings on a cumulative reduction of output and income until saving and investment are brought down to equality with the lesser investment demand. Insufficiency of investment demand has been identified with depression and recessions and tendencies towards chronic unemployment. Stimuli to investment, such as reductions in tax rates on income from capital, have thus seemed in order to bring investment demand up to the levels of saving that would be forthcoming with full employment. Conversely, excessive levels of investment demand can create inflationary pressures, calling for policies that would restrict investment.

These Keynesian perceptions as to the costs and benefits of investment are startlingly different from those of the classical models – old and new – which assume, implicitly or explicitly, that the economy is operating at full employment and full utilization of resources. In the classical models, more current production of capital must mean less current production of consumption goods and services. And more consumption now must mean less current investment and less output and consumption in the future.

In an economy with substantial unemployed resources, however, more investment need not and probably will not bring less consumption. Expenditures for additional investment will rather constitute additional incomes for their recipients, and this income will in turn largely be spent on increased consumption. Thus the production of consumption goods and services will increase rather than decline. And more consumption may bring about more investment, as producers see a need for additional capital to increase the output of consumption goods and services.

Classical and Keynesian views also differ on the principal mechanism by which intended investment and saving are equated. In the classical view, changes in the rate of interest are presumed to perform this task. Investment demand is thought to be negatively related and very sensitive to the rate of interest, which is the cost of borrowing funds to finance capital spending. If investment demand is smaller than saving at the full-employment level of incomes, the classical analysis holds that the excess of funds in the credit market will depress interest rates, thereby inducing increases in investment demand (and possibly reductions in saving as the interest earned by savers falls) until intended investment and saving

are equal. Thus, no change in the level of economic activity (output) need occur as in the Keynesian analysis. The Keynesian view of the equilibrating process has interest rates playing a smaller role than changes in output, because investment demand is thought to be relatively insensitive to interest rates, being dominated instead by producers' expectations of future demand for their products. Even if the investment demand were sensitive to interest rates, expectations could be so pessimistic that, even if the rate of interest were to fall to zero, there would be insufficient investment demand.

Empirical studies have attempted to measure the influence of interest rates, taxes and expectations of future demand on investment decisions. Producers are presumed to acquire capital to increase their expected profits. The profitability of additional capital depends on its cost, on its expected productivity and on expectations on the price at which additional output can be sold. On the assumption that output is a fixed, 'well-behaved' function of capital and labour (strictly concave, with declining partial derivatives of output with the respect to capital and labour and positive cross-partial derivatives), producers will acquire capital to the point where its declining marginal product equals its cost. This will then define both the desired, or equilibrium, capital–labour ratio and capital–output ratio. With the supply of labour and the rate of output fixed and no change in the relative price of capital and labour, investment in equilibrium will be equal to depreciation, or what is necessary to maintain the existing capital stock, and net investment will be zero. Positive net investment will then stem from increases in the demand for output or reductions in the relative price of capital. Increases in output will generate investment demand to maintain the equilibrium capital–output ratio. A reduction in the cost of capital would generate investment in order to increase the capital–labour and capital–output ratios. In either case, maintaining increased amounts of capital will generate further investment to cover increased depreciation.

In general, the desired capital stock may be written as:

$$K^* = f(p, c, Y^*), \tag{1}$$

where p is the price of output, $c = q[i - (\dot{q}/q) + d]$ is the rental price or user cost of capital, q is the supply price of capital goods, i is the opportunity cost of capital, d is the rate of economic depreciation and Y^* is desired output. If firms minimize expected costs of producing an exogenously given or expected output Y, then the wage rate, w, would be substituted for p. The rental price, or user cost of capital, c, is the cost per period of holding and maintaining one unit of capital. In the absence of taxes, it is the price of capital goods multiplied by the sum of the real interest rate and the rate of economic depreciation. The former measures the opportunity cost in terms of forgone net earnings from lending or otherwise investing money, plus the capital loss (or minus the capital gain) associated with changing prices of capital goods.

Building on this neoclassical theory of the firm developed by Haavelmo (1960), and assuming a Cobb–Douglas production function with elasticity of output with respect to capital, b, Jorgenson (1963, 1967) arrived at a demand function for capital with a particular form that has been employed in a large number of influential studies:

$$K^* = b(p/c)Y^*. \tag{2}$$

With an implicit unitary elasticity of K^* with respect to c, this formulation implies strong effects of monetary policy, via the rate of interest, and of tax policy so far as, by accelerated tax depreciation, investment subsidies or exclusion of capital gains from taxation, it affects the value of c (see below).

The more general constant-elasticity-of-substitution (CES) production function may be used to generate a demand for capital having the form:

$$K^* = h(p/c)^s (Y^*)^r, \tag{3}$$

where s, the elasticity of substitution between labour and capital, is the critical elasticity of demand for capital with respect to the relative price of capital, and r is the elasticity of demand for capital with respect to output. The elasticity, r,

will be greater than, equal to, or less than unity as the returns to scale are decreasing, constant or increasing.

If relative prices are constant, or if technology requires that capital and labour be used in fixed proportions (in which case the elasticity of substitution is zero), then with constant returns to scale, desired capital is proportional to the demand for output. This form of the demand for capital leads to the 'acceleration principle', according to which net investment demand, arising from a desire to change the stock of capital, depends not on the level of demand for output, but on the *change* in demand for output (Clark 1917). To induce firms to invest (acquire more capital), demand for output must be expected to rise. Both the original formulation by Jorgenson of the demand for capital (2) and the more general formulation (3) underly *a* 'flexible accelerator', where the desired capital–output ratio is not constant but depends on prices and on the scale of output and, as seen below, investment is subject to a distributed lag process (Koyck 1954) affected by adjustment costs and the dynamic process governing the formation of expectations of future variables (Eisner and Strotz 1963; Helliwell and Glorieux 1970; Lucas 1976; Eisner 1978).

Many early econometric studies of investment behaviour tested the accelerator in various forms, but generally they did not allow for effects of prices on the desired capital–output ratio, which is the hallmark of Jorgenson's neoclassical approach. The major competing hypothesis was that investment depends on the level of profits, on the grounds that realized profits measure expected profits, or that capital market imperfections cause firms' capital expenditures to be constrained by the flow of internal funds (Meyer and Kuh 1957). Reviews of these earlier investigations are found in Eisner and Strotz (1963) and Jorgenson (1971). The practice in recent studies has been to capture profit expectations by including expectations of the major determinants of profits, namely, sales, prices and wages, or to approximate them by stock market valuations of firms. The flow of internal funds may play some role in investment decisions, not as a determinant of the desired capital stock but as a factor influencing the speed of adjustment of capital (Coen 1971).

To study the effects of tax policy on demand for capital, the rental price can be generalized to incorporate parameters of the tax system. For example, the after-tax cost of holding one unit of capital would be:

$$c = q\left[(1-uv)i-(1-uw)\left(q^{j}/q\right)+d\right]\times[1-k-uz]/(1-u) \tag{4}$$

where u is the rate of taxation of business income; v is the proportion of the opportunity cost of capital (such as interest, dividends and forgone earnings) that is tax deductible; w is the proportion of capital gains and losses effectively taxed; k is the effective rate of the investment tax credit or subsidy; and z is the present value of the tax depreciation expected from a dollar of investment (Hall and Jorgenson 1967).

It can be seen, in this definition, that higher values of v, k and z (from accelerating tax depreciation) reduce the value of c, as does a higher value of w, provided that capital goods prices are expected to rise. The value of c would also be lowered by decreasing the rate of interest or other measure of the opportunity cost of capital. A higher rate of inflation of capital goods prices has two opposing effects on c. In so far as higher inflation reduces the real after-tax opportunity cost of capital, it reduces c. However, if tax depreciation is based on the historical cost of assets rather than on replacement costs, inflation reduces the present value of tax allowances, z, and thereby raises c (Feldstein 1982). Finally, we may note that changes in the general rate of business taxation are ambiguous in their effects on c. If v and w are unity, and if the opportunity cost of capital is unaffected by a change in the business tax rate, then a decrease in u will reduce, leave unchanged or increase c as the present value of tax allowances on a unit of investment (including the investment credit) is less than, equal to or greater than the present value of economic depreciation (Hall and Jorgenson 1971). But then, going back to Eq. 1, the effect of any of these parameters on K^* depends upon the elasticity of the latter with respect to c.

The desired capital stock does not in itself indicate the rate of investment, which is the rate of replacement of existing capital plus the rate of net additions. Both entail a combination of financial considerations and costs of adjustment, which will in turn relate to costs of acquiring information necessary to decisions, costs of planning and the supply function for capital goods, all filtered through the expectations of agents.

If adjustment costs are an increasing function of the rate of investment, it will generally prove optimal not to adjust capital to the desired level immediately, but instead to distribute changes in the capital stock over time (Eisner and Strotz 1963). The speed of adjustment of capital to changes in its desired or equilibrium level may depend on the causes and magnitudes of the changes. An increase in the demand for output may generate investment with all due speed as expectations become firm with regard to the permanence of the increased demand. If, however, the increased demand for capital is due to a fall in its relative price (because, let us say, of a reduction in the rate of interest), thus generating a demand for more durable and hence more substantial and expensive capital, the rate of investment may be slowed by the availability of existing capacity sufficient for current production. These considerations underlie the 'putty-clay' model in which the capital–labour ratio can be varied on newly installed capacity but cannot be altered on existing capacity. A demand for additional housing services will bring on investment in housing as rapidly as cost considerations permit. A lower rate of interest, causing substantial investment in more durable brick houses to replace less durable houses of wood or straw, would cause the rate of investment to increase only as existing houses of wood and straw wear out and are replaced.

Investment equations should thus in principle involve separate distributed lag responses to changes in relative prices and to changes in output. They should also admit the possibility that the lag distribution is not fixed and may vary with other economic parameters and the expectations function.

A logarithmic transformation of Eq. 3 yields

$$\ln K^* = \ln h + \mathrm{s} \ln (p/c) + \mathrm{r} \ln Y. \tag{5}$$

Putting this in first difference form, we have:

$$\Delta \ln K^* = s\, \Delta \ln(p/c) + r\, \Delta \ln Y. \tag{6}$$

Since the change in the logarithm of capital is the relative change in capital, we may treat the ratio of net investment to existing capital stock as approximately equal to $\Delta \ln K$, which may in turn be written as a distributed lag function of changes in the determinants of desired capital:

$$I_N/K_{-1} = s[q_1(L)\, \Delta \ln(p/c)] + r[q_2(L)\, \Delta \ln Y], \tag{7}$$

where $q_1(L)$ and $q_2(L)$ are lag operators that indeed should be functions of such variables as the rate of interest, and the cost and availability of capital. Then, finally, since investments equal net investment plus replacement, we may write

$$I = I_N + R = I_N + dK_{-1}, \tag{8}$$

where d, the replacement rate, may vary over time.

Estimates of investment functions of this type have often neglected influences of economic variables and expectations on adjustment processes and the replacement rate. Lag distributions are assumed to be of some fixed functional form, and d is assumed to be constant (for evidence that d may not be constant, see Feldstein and Foote 1971; Eisner 1972; Feldstein and Rothschild 1974; Coen 1975).

Where production and lag parameters have not been unduly constrained by a priori specifications, estimates have generally yielded values of s, the elasticity of substitution, considerably less than unity, in some cases not substantially greater than zero (see Eisner and Nadiri 1968; Coen 1969; Lucas 1969; Eisner 1978; Chirinko and Eisner 1982). Lag distributions estimated from time series and cross-section data have usually extended over a number of years (Eisner 1978), and they often have inverted-U shapes. Where a putty-clay formulation has been employed with separate lags on relative prices and output, the

mean lags on prices are typically much longer than those on output (Bischoff 1971).

These findings of small price elasticities of demand for capital suggest some role, but a limited one, for monetary and tax policies in directly affecting the general rate of investment through the rental cost of capital. The long lag distributions on relative prices suggest further difficulties in the use of monetary or fiscal policy for reducing cyclical fluctuations in investment. However, policy impacts may operate not only on the desired capital stock but also on the speed of adjustment of capital.

Repeated changes in tax parameters such as k, the rate of investment tax credit or subsidy, may be used to bring about intertemporal substitution of investment even if the effects on its long-run average are small. Thus, when investment is low, the marginal rate of subsidy, k, might be raised, while if investment were deemed too great, the value of k could be reduced to zero or indeed made negative (an investment tax instead of subsidy). Paradoxically, a fluctuating and uncertain investment subsidy/tax may have substantial effects on investment where permanent subsidies or taxes would not. There is thus an asymmetry between effects of changes in the cost of capital and changes in the demand for output, the effects of which on investment will be proportional to the permanence with which they are perceived (Eisner 1978).

Most investment functions, with their ad hoc, fixed lag distributions and assumptions of static expectations, fail to capture accurately the effects of economic policies on the timing of investment or to distinguish properly between the effects of temporary and permanent policy changes (Lucas 1976). To correct these shortcomings, adjustment costs must be explicitly introduced in the firm's optimization problem, so that instead of there being a desired stock of capital towards which the firm moves in a mechanistic way, there is a desired path of capital accumulation. Along such a path, the optimal rate of investment at each point, including the present, will in general depend on expected relative prices and output over the entire planning horizon.

Obtaining solutions to the firm's dynamic optimization problem under very general specifications of technology and expectations has proven difficult. To make such an approach empirically tractable, strong assumptions are usually made, for example, that the production function is quadratic, that adjustment costs are quadratic, symmetric and separable from the rest of technology (the cost of adjusting capital, for example, does not depend on the quantities of capital and labour currently employed), and that expectations are characterized by relatively simple autoregressive processes.

The critical role in current investment of unobservable adjustment costs and of uncertain, shifting (and generally not directly observable) expectations of the future, stressed by Keynes, has sparked interest in a formulation of an investment function that directly relates demand prices and supply prices of capital. Going back to Keynes's *General Theory,* we have investment undertaken to the point where the expectation of marginal profit on investment (the 'marginal efficiency of investment') is equal to the rate of interest or, alternatively, the present value of expected returns from the marginal investment, using the rate of interest as the discount factor, is equal to the marginal supply price of newly produced capital goods. Building on this, Brainard and Tobin (1968) and Tobin (1969), presented a 'q-theory', which sees investment as a positive function of the ratio, q, of the market value of capital to its replacement cost. The former may in principle be observed in the trading prices of stock shares along with bonded indebtedness of business firms. With proper adjustment for tax considerations, when the value of q is greater than unity, investment will take place because the cost of additional capital will be less than the market evaluation of the present value of returns from capital. Conversely, when q is less than unity, business demand for capital may be better satisfied by acquisitions taking over existing firms and their facilities than by new investment. In general, the rate of investment should be greater the greater the value of q.

Empirical estimation of 'q' investment equations and predictions based on these estimates have not, however, proved very successful (von Furstenberg 1977; Abel 1980; Summers 1981;

Hayashi 1982; Abel and Blanchard 1986). Suggested explanations of the difficulties include the fact that market values of firms may relate to much more than the tangible capital generally included in business investment, and the failure to distinguish marginal and average values of the cost of new capital versus the acquisition costs of existing firms (Chirinko 1986).

Investment decisions are but one element of producers' plans for hiring or acquiring factors of production. Interrelationships between investment demand and demands for other inputs have been a subject of growing interest. Since factor demands are derived from a given production function, they share common technological parameters and may be estimated as a system of demand functions (Coen and Hickman 1970). Such an approach calls attention to effects of investment stimuli that are often overlooked. For example, at a given level of output, the direct impact of an investment tax credit is to reduce the demand for labour, since it raises the relative cost of labour. Employment may eventually be raised, but only if the expansion in aggregate output induced by the increase in investment demand is large enough to offset the adjustment to a higher capital–labour ratio.

Additional interrelationships may arise when capital is not the only factor of production subject to adjustment costs. If labour input is also costly to change, then the rate of investment may depend not only on the desired adjustment in capital stock but also on the desired adjustment in employment (Nadiri and Rosen 1969; Brechling 1975; Epstein and Denny 1983). Furthermore, since a firm must operate on its production function, factor adjustments cannot be entirely independent. If output is exogenously given and there are n inputs, $n - 1$ inputs can be independently adjusted, but the nth is determined by the production function, the level of output, and the quantities of the other inputs (Gould 1969). It may be unreasonable to view the production function as a binding constraint, however, because it is difficult, if not impossible, to measure perfectly all inputs and their utilization rates.

With the development of dynamic optimization models of interrelated factor demands in which various types of capital and other inputs are subject to adjustment costs, and expectations are not treated as static, it is possible to estimate the magnitude of adjustment costs for capital, to see how they affect and are affected by adjustments of other inputs, and to study the impacts of changes in producers' perceptions of the *processes* generating prices, output and policy parameters. As we noted above, this approach necessitates strong restrictions on functional forms to obtain explicit decision rules for accumulation of capital and employment of other inputs (Meese 1980). Where general forms of the production, adjustment cost and expectations functions are assumed and the model cannot be solved completely, it is still possible to estimate the first-order conditions (Euler equations) that implicitly define the evolution of the optimal inputs (Pindyck and Rotemberg 1983; Shapiro 1986). Such estimates do not give a complete account of the dynamics of investment behaviour for any initial conditions and stochastic environment, but they do give insights about differing short- and long-run responses to, say, an unexpected increase in the price of energy starting today versus the same increase beginning 5 years from now but anticipated today. An important empirical development in the study of investment is the use of more disaggregated data-sets; an important example is Cummins et al. (1994), which analyses the effects of major tax reforms on investment based on firm-level panel data. The paper is important in providing much stronger evidence on the importance of the user cost of capital than appears in aggregate studies. Chirinko (1993) is still of value as a survey.

Yet, as noted in the valuable review of Caballero (1999), a general dissatisfaction with the empirical performance of the neoclassical model led to change in investment research which emphasized the role of irreversible. See irreversible investment for these new developments.

See Also

► Irreversible Investment
► Neoclassical Synthesis

- ▶ New Classical Macroeconomics
- ▶ Public Capital
- ▶ Rational Expectations
- ▶ Tobin's Q

Bibliography

Abel, A.B. 1980. Empirical investment equations: an integrative framework. In *On the state of macroeconomics*, ed. K. Brunner and A.H. Meltzer. Vol. 12 of Carnegie-Rochester Conference on Public Policy. Amsterdam: North-Holland.

Abel, A.B., and O.J. Blanchard. 1986. The present value of profits and cyclical movements in investments. *Econometrica* 54: 249–273.

Bischoff, C.W. 1971. The effect of alternative lag distributions. In *Tax incentives and capital spending*, ed. G. Fromm. Washington, DC: Brookings Institution.

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

Brechling, F. 1975. *Investment and employment decisions*. Manchester: Manchester University Press.

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

Chirinko, R.S. 1986. Business investment and tax policy: A perspective on existing models and empirical results. *National Tax Journal* 39: 137–155.

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

Chirinko, R.S., and R. Eisner. 1982. The effects of tax parameters in the investment equations in macroeconomic econometric models. In *Economic activity and finance*, ed. M.E. Blume, J. Crockett, and P. Taubman. Cambridge, MA: Ballinger.

Clark, J.M. 1917. Business acceleration and the law of demand: A technical factor in economic cycles. *Journal of Political Economy* 25: 217–235. Reprinted in American Economic Association, Readings in Business Cycle Theory. Philadelphia: Blakiston, 1951.

Coen, R.M. 1969. Tax policy and investment behaviour: Comment. *American Economic Review* 59: 370–377.

Coen, R.M. 1971. The effect of cash flow on the speed of adjustment. In *Tax incentives and capital spending*, ed. G. Fromm. Washington, DC: Brookings Institution.

Coen, R.M. 1975. Investment behavior, the measurement of depreciation, and tax policy. *American Economic Review* 65: 59–74.

Coen, R.M., and B.G. Hickman. 1970. Constrained joint estimation of factor demand and production functions. *The Review of Economics and Statistics* 52: 287–300.

Cummins, J.G., K.A. Hassett, and R.G. Hubbard. 1994. A reconsideration of investment behavior using tax reforms as natural experiments. *Brookings Papers on Economic Activity* 1994 (2): 1–59.

Eisner, R. 1972. Components of capital expenditures: Replacement and modernization versus expansion. *The Review of Economics and Statistics* 54: 297–305.

Eisner, R. 1978. *Factors in business investment*. Cambridge, MA: Ballinger.

Eisner, R. 1985. The total incomes system of accounts. *Survey of Current Business* 65: 24–48.

Eisner, R. 1986. *How real is the federal deficit?* New York: Free Press, Macmillan.

Eisner, R., and M.I. Nadiri. 1968. Investment behavior and the neo classical theory. *The Review of Economics and Statistics* 50: 369–382.

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

Epstein, L.G., and M.G.S. Denny. 1983. The multivariate flexible accelerator model: Its empirical restrictions and an application to U.S. manufacturing. *Econometrica* 51: 647–674.

Feldstein, M.S. 1982. Inflation, tax rules and investment: Some econometric evidence. *Econometrica* 50: 825–862.

Feldstein, M.S., and D.K. Foote. 1971. The other half of gross investment: Replacement and modernization expenditures. *The Review of Economics and Statistics* 53: 49–58.

Feldstein, M.S., and M. Rothschild. 1974. Towards an economic theory of replacement investment. *Econometrica* 42: 393–423.

Gould, J.P. 1969. The use of endogenous variables in dynamic models of investment. *Quarterly Journal of Economics* 83: 580–599.

Haavelmo, T. 1960. *A study in the theory of investment*. Chicago: University of Chicago Press.

Hall, R.E., and D.W. Jorgenson. 1967. Tax policy and investment behavior. *American Economic Review* 58: 391–414.

Hall, R.E., and D.W. Jorgenson. 1971. Application of the theory of optimal capital accumulation. In *Tax incentives and capital spending*, ed. G. Fromm. Washington, DC: Brookings Institution.

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

Helliwell, J.F., and G. Glorieux. 1970. Forward looking investment behavior. *Review of Economic Studies* 37: 499–516.

Hicks, J.R. 1950. *A contribution to the theory of the trade cycle*. Oxford: Clarendon Press.

Jorgenson, D.W. 1963. Capital theory and investment behavior. *American Economic Review: Papers and Proceedings* 53: 247–259.

Jorgenson, D.W. 1967. The theory of investment behavior. In *Determinants of investment behavior*, ed. R. Ferber. New York: Columbia University Press.

Jorgenson, D.W. 1971. Econometric studies of investment behavior: A survey. *Journal of Economic Literature* 9: 1111–1147.

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

King, M.A., and D.K. Fullerton. 1984. *The taxation of income from capital*. Chicago: University of Chicago Press.
Koyck, L.M. 1954. *Distributed lags and investment analysis*. Amsterdam: North-Holland.
Lange, O. 1938. The rate of interest and the optimum propensity to consume. *Economica* 5: 12–32. Reprinted in American Economic Association, Readings in Business Cycle Theory, Philadelphia: Blakiston, 1951.
Lucas, R.E. 1969. Labor-capital substitution in U.S. manufacturing. In *The taxation of income from capital*, ed. A.C. Harberger and M.J. Bailey. Washington, DC: Brookings Institution.
Lucas, R.E. 1976. Econometric policy evaluation: A critique. In *The Phillips curve and labor markets*, ed. K. Brunner and A.H. Meltzer. Vol. 1 of Carnegie-Rochester conference in public policy. Amsterdam: North-Holland.
Meese, R. 1980. Dynamic factor demand schedules for labor and capital under rational expectations. *Journal of Econometrics* 14: 141–158.
Meyer, J., and E. Kuh. 1957. *The investment decision*. Cambridge, MA: Harvard University Press.
Nadiri, M.I., and S. Rosen. 1969. Interrelated factor demand functions. *American Economic Review* 59: 457–471.
Pindyck, R.S., and J.J. Rotemberg. 1983. Dynamic factor demands and the effects of energy price shocks. *American Economic Review* 73: 1066–1079.
Shapiro, M.D. 1986. The dynamic demand for capital and labor. *Quarterly Journal of Economics* 101: 513–542.
Summers, L.H. 1981. Taxation and corporate investment: A *q*-theory approach. *Brookings Papers on Economic Activity* 1981 (1): 67–140.
Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1: 15–29.
von Furstenberg, G.M. 1977. Corporate investment: Does market valuation matter in the aggregate? *Brookings Papers on Economic Activity* 1977 (2): 347–397.

Investment and Accumulation

Stephen A. Marglin

The standard view of accumulation goes something like this. In the short period, fraught with frictions and maladjustments, the demand for investment interacts with the supply of saving, more or less *à la* Keynes, to determine the growth of the capital stock. Keynesian policies may have fallen into disrepute, but for the short period the representative economist continues to use the tool box developed in the *General Theory* and its wake: accumulation falls out from the determination of national income.

In the longer period the same economist falls back on very different arguments: the mainstream of the profession takes accumulation to be determined by saving propensities, with nary a side glance at investment demand. That Japan has over the last quarter century devoted 30 per cent of gross output to fixed capital formation and Great Britain 20 per cent is conventionally explained in terms of higher Japanese saving propensities, not in terms of a greater propensity to invest.

In a still longer time frame, even saving propensities become irrelevant. In the asymptotic future beyond all future, accumulation is determined solely by population growth and technical change. Saving propensities may affect the steady-state capital: output ratio if the technology admits of substitution between labour and capital, but that is the limit of their influence.

Economists of a Marxian bent share the mainstream view, up to the asymptotic future, which they rightly dismiss as an irrelevant construct. The terminology may differ: difficulties of 'realization' is favoured for describing a shortfall of aggregate demand relative to aggregate supply, and hence (abstracting from foreign trade and government surpluses or deficits) for a shortfall of investment relative to saving. But for Marxians realization problems are generally confined to the short period; in the long run it is once again the saving propensities of capitalists, along with the rate of profit determined by class struggle, which determine the rate of accumulation. To be sure, neoclassical and Marxian theories of the determination of saving propensities differ, but for present purposes this is a secondary issue; the short run apart, the two theories agree that investment propensities are irrelevant to accumulation.

Investment in the General Theory

Against these views stands the Keynesian view, which, applied to the long run, tells a very

different story of the accumulation process. In the *General Theory* Keynes formalized his view of investment demand in terms of a 'marginal efficiency of capital' schedule which showed the amounts of investment that would be forthcoming at different rates of interest. The basic idea behind this schedule can be captured by supposing there to be a set of projects indexed by $i = 1, \ldots, n$, each requiring a unit of investment and returning, respectively, a cash flow of $r_1, \ldots, r_n$ in perpetuity. If the n projects are arrayed in descending order of r_i, then with the simplifying assumption that each investment costs one dollar, r_i is the *marginal* rate of return on the investment of i dollars. In Keynes's language, r_i is the marginal efficiency of capital.

Suppose now that the interest rate d, which represents the cost of capital to the investor, is also expected to be constant in perpetuity. Then the present value of the ith project's return is r_i/d. The profit-maximizing investor will go down the list until he reaches the project at which $r_i = d$, that is, the point at which the marginal rate of return just equals the cost of capital. In Keynesian terms, investment is determined by equating the marginal efficiency of capital to the cost of capital. More precisely, the array of projects being discrete, the profit-maximizing rule is to undertake all projects for which $r_i > d$ and to reject those for which $r_i < d$. If there is a project for which the relationship between r_i and d is one of exact equality, it is a matter of indifference whether the project is undertaken or not. The main point is that the discounted present values of returns, $r_i/d - d$ acts as the discount rate – exceed the assumed unit cost of the investment provided $r_i > d$.

So far there is nothing novel in this theory. Knut Wicksell would have had no problem making the argument his own, and indeed the schedule of the marginal efficiency of capital bears a close resemblance to the marginal productivity of capital schedule of mainstream theories of accumulation. Even the overall theoretical structure which Keynes builds by joining this schedule to schedules of consumption and liquidity preference would not have been uncongenial to Wicksell, particularly if its application is confined to the short period. This is presumably why Wicksell's Swedish followers chided Keynes for indulging 'the attractive Anglo-Saxon kind of unnecessary originality' (Myrdal 1939, p. 8; the comment naturally refers to the *Treatise*, not to the *General Theory*).

Probably because of the affinity to a Wicksellian version of neoclassical economics, the mainstream of American economists has been able, as was suggested at the outset of this essay, to accept a version of the Keynesian analysis for the short period – much to the annoyance of Keynes's Cambridge disciples, like Joan Robinson and Nicholas Kaldor, who all their lives insisted that Keynes's main message was being lost in the translation. In the standard American view, the main point of Keynes was the limit of monetary policy, conceived of in terms of its affect on d, to affect investment demand, either because of a low elasticity of the marginal efficiency schedule, or, in the limit, because of the impossibility of reducing the interest rate (the famous 'liquidity trap', of which Keynes said in the *General Theory* that it was, as yet (1935), a theoretical possibility of which there had been no actual instances). The bottom line was the need not only for state intervention in the form of an activist monetary policy – this was fully present in Wicksell's analysis – but also in the form of fiscal policy. Indeed it does no disrespect to Keynes to accept that his influence owed as much to the intellectual justification he provided for an activist, interventionist state – for the end to *laissez faire* – as to the intellectual power of his ideas, certainly as these ideas were, reflected through the prism of the mainstream of the American economics profession.

An Alternative Reading

But there is another reading of Keynes. The real departure of the Keynesian theory of investment from the orthodox one, in my judgment, starts from a recognition that the formalism of his theory of investment demand obscures its real content. The starting point is the recognition that the returns of any project, lying in the future, are inherently uncertain. The r_i's are not objectively

given reality but a subjective construction of the investor. It has become fashionable to blur the old Knightian distinction between risk, objective and quantifiable, and uncertainty, subjective and qualitative, by means of the theory of subjective or personal probability. Even if the axioms underlying this theory are neither compelling (particularly the assumption of a complete ordering and the assumption of 'independence', what Leonard J. Savage called the 'sure thing' principle), nor borne out empirically in the behaviour of untutored individuals, subjective probability theory still has some heuristic value in modelling investment decisions, particularly in its emphasis on the psychology of the decision maker.

Subjective probability allows us to go behind the r_i's of present value calculations like the simple perpetuity formula r_i/d to more complex sums of the form

$$r_i^h = p_1^h u_1^h r_{i1}^h + p_2^h u_2^h r_{i2}^h + \cdots + p_m^h u_m^h r_{im}^h,$$

in which the generic term $p_j^h u_j^h r_{ij}^h$ is composed of these elements: p_j^h is Mr h's subjective probability of the occurrence of a particular complex of events (a 'state') in which his marginal utility of income (normalized) is u_j^h and his estimated return from project i is r_{ij}^h. The central point is that in the Keynesian view, each of the constituents – the probability p_j^h, the marginal utility of income u_j^h, and the state-specific return r_{ij}^h – owes as much to the imagination of the investor as it does to an objective reality. The more optimistic are investors, the higher the probabilities they will attach to states in which the returns and the marginal utilities of these returns are high, and the consequence will be higher values of r_i^h. Conversely, the more pessimistic are investors, the lower the r_i^h's that will be attached to the same projects. Thus the 'animal spirits' of investors play a crucial role in investment demand.

The recognition of a crucial role for animal spirits directs one's attention away from movements *along* the marginal efficiency schedule. The question becomes, what determines the position of this schedule? Evidently, according to what has just been said, it is, in the last instance, investors' evaluations of probabilities of various states, of the relative utility of income in different states, and of returns in different states.

It is equally evident that this makes a cumbersome theory. A more tractable model can be constructed by making the prospective returns the r_i^h's, depend on the general anticipations of capitalists. The higher is the general expectation of profits, the higher will be the anticipated rate of return on specific projects – a rising tide will be anticipated rate of return in specific projects – a rising tide will presumably lift all boats. In this view, movements of the entire marginal efficiency schedule, triggered by changes in expectations of profitability, not movements along a given schedule induced by changes in the interest rate, are the key to understanding the ups and downs of investment and output.

Such reasoning – this is of course a 'rational reconstruction' – permitted Keynes's heirs, Roy Harrod in Oxford and Robinson and Kaldor in Cambridge, who re-situated the *General Theory* in a long-period context, to recast the marginal efficiency schedule in terms of variables related to expected profits. Robinson's investment demand function, for example, and the argument of Keynes's own formulation, the rate of interest, disappears into the background of *ceteris paribus*.

The rationale for ignoring the cost of capital is the assumption of a highly insensitive responsiveness of investment and saving to plausible interest rate changes. Clearly, there are strong assumptions at work here about the relevant range of interest rate variation. As long as the anticipated returns are finite, there must be *some* level of the interest rate at which even the most attractive projects appear to all and sundry as uneconomic. Thus, unless saving is negatively related to the rate of interest *and* highly elastic, sufficient variation in interest rates could, with enough time, adjust the demand for investment to the supply of saving.

For most of the postwar period, however, the range of variation of interest rates has been too modest to test this possibility, at least if we identify the interest rate with the difference between the nominal rate on government or high grade corporate bonds and the rate of inflation. Indeed, from 1945 until 1980 this 'real' rate of interest

never moved very far from zero in the United States. In the post-1980 disinflation and recession, real interest rates rose to levels which must give pause to even the most devoted neo-Keynesian. It is certainly too early at this writing to determine whether in the sweep of history this was a momentary aberration or the dawn of a new era; my own leaning is towards aberration, but that may reveal my neo-Keynesian predilections rather than a reasoned guess about the future.

The rate of profit generally expected on the capital stock as a whole (*re*) is itself no more observable than any other variable that lies in the future. In the neo-Keynesian literature *re*'s customarily are taken to be a function of a small number of variables which summarize the relevant information available to investors. The standard version of the theory simply extrapolates the current or immediate past rate of profit.

With simplifying assumptions like constant returns to scale, homogeneous capital, and a uniform, exogenously given rate of capacity utilization, it must be true *ex post* that the average rate of profit on new investment (which is the marginal rate of profit on the capital stock) turns out to equal the average rate of profit on all capital. Under these assumptions it might appear reasonable for the expected average rate of profit on the entire capital stock *re* to equal the expected rate of profit on new investment, which would impose, in the spirit of 'rational expectations', a consistency requirement for each investor

$$r^e \equiv \sum_i r_i^h \equiv \sum_i \sum_j p_j^h u_j^h r_{ij}^h.$$

But no such inference is warranted. Neo-Keynesian theory in my view is compatible with rational expectations – insofar as this notion makes any sense at all under conditions of subjective uncertainty. But an important measure of realism would be lost by constraining the theory in this way, for even under the stringent assumptions that lead to a uniform realized profit rate', there is no good reason to believe in rational expectations. The general expectation of a 4 per cent return is perfectly consistent with individual expectations that special gifts, opportunities, or kismet will result in a 10, 20 or even 50 per cent return on one's own projects. As P.T. Barnum would have it, a sucker is born every minute.

Self-Fulfilling Prophecy

Keynes made Barnum's sucker into a kind of hero: in Keynes's view it was doubtful that capitalism could function without such a 'spontaneous urge to action'. Neo-Keynesian theory makes an even stronger assertion, which has its roots in Keynes's *Treatise on Money*. As long as there are enough of them, the 'suckers' can turn the tables on the rest of us. Capitalists, as a class, have the power to shape conditions so that their expectations come true, at least in large enough part to maintain their confidence.

Unique among economic actors, this class has the power of self-fulfilling prophecy. Not only do actual profit rates affect capitalists' beliefs about future profits, capitalists' beliefs also have an impact on actual profits. This is not the same thing as the ability Joan Robinson imputed to capitalists to make the profit rate anything they liked, but it is a formidable power nonetheless.

There are two mechanisms by which capitalists' prophesies about profits become self-fulfilling, the distribution of income between capital and labour and the level of capacity utilization and employment. The first is the less familiar one, except to readers of Keynes's *Treatise on Money*. Suppose a closed economy with no government spending or taxation in which the starting point is a long run equilibrium characterized by equality both between desired saving and investment and between expected and actual (average) rates of profit. Imagine a change in 'animal spirits' which makes capitalists willing to undertake more investment at the going rate of profit than earlier was the case. In other words, imagine an outward shift in the investment demand function. In the simplest neo-Keynesian story, this addition to aggregate demand increases spending relative to income – remember *income* has remained unchanged – and drives the price level upward. Assuming money wages are fixed or at least sluggish, this reduces the real wage and shifts the

income distribution in favour of profits. The process continues since higher realized profits lead to further expectations of higher profits and still more investment demand. But it does not continue indefinitely, because capitalists are assumed to save a higher proportion of their incomes than do workers. The upward spiral of prices and investment continues only until the extra saving induced by higher profits absorbs the extra investment, at which point the economy comes to a new equilibrium where desired saving and investment and actual and anticipated profits are agains equal, albeit at a higher level. In addition to the existence of an investment demand function that is distinct from the saving function, the sluggishness of money and wages and the difference in saving propensities between classes are crucial to this result.

Several observations are in order here. First, although the *fons et origo* of this theory is Keynes's *Treatise*, the theory has much in common with the model outlined in Josef Schumpeter's *Theory of Economic Development*. There are two important conceptual differences however. First, the neo-Keynesian equilibrium is formulated as a steady growth rather than stationary (zero growth) state that dominated in Schumpeter's time. Thus shifts from one equilibrium to another involve changes in the rate of growth rather than changes in the level of output.

A second difference is more fundamental. In both the Schumpeterian and the neo-Keynesian view, an outward shift in the investment demand function plausibly involves an expansion of the array of projects yielding returns in excess of the cost of capital. And in both views, the psychology of the capitalist class is crucial. Schumpeter no less than the neo-Keynesians recognized the subjective element in the estimation of returns. 'Invention', for Schumpeter, was necessary but not sufficient for 'innovation'. But here the resemblance ends. In the neo-Keynesian view invention is neither necessary nor sufficient for innovation. Investment can be a pure boot-strap operation; the theory requires nothing more than a change in business psychology to change investment demand, and a change in investment demand can lead the economy to a new equilibrium with a different rate of growth. Within the limits of saving propensities and the malleability of real wages, capitalists wishes are self-fulfilling. Prices and profit rates change to validate the changes in capitalists' expectations!

The Crucial Assumptions

There must be a trick. In fact, there are four crucial assumptions, two of which – the flexibility of real wages and the difference in propensities to save between capitalists and workers – have already been mentioned. The assumption that wages are set in money terms plays a central role in the analysis. Money wages need not be fixed once and for all, but Robinson's version of the neo-Keynesian story (and the Schumpeterian story for that matter) cannot be told at all without the assumption of sluggish money wages. Evidently it is the income distribution that adjusts saving and investment to each other. So if the income distribution is fixed, then it cannot do the job which neo-Keynesian theory assigns it.

A difference in saving propensities is equally important to the neo-Keynesian view of capitalism. There are various versions of the so-called Cambridge saving equation, and a fair amount of confusion about the content of alternative versions exists two decades after the most important contributions to this debate. But all versions of the theory take it for granted that capitalists' propensity to save exceeds workers'. By contrast, the principal neoclassical theory of saving, Franco Modigliani's life-cycle hypothesis, suggests that the propensity to save out of wages will exceed the propensity to save out of property income: wages will be disproportionately in the hands of people preparing for retirement and profits disproportionately in the hands of retirees. Theoretical dispute is of course nothing new in economics. What is more surprising is that we lack persuasive empirical evidence of one view or the other two decades after the theoretical battle was fairly joined.

The essential role that flexible real wages and differences in saving propensities play is relatively transparent, so perhaps 'trick' is not an

appropriate description for either of these assumptions. But a third assumption is necessary to make capitalists' investment spending into a 'widow's cruse' (Keynes's metaphor), filling up with saving as fast as it is emptied in investment. This assumption is better hidden. We can see its role more clearly by asking how the process described for moving from one equilibrium in consequence of a shift in investment demand ever gets started. How is it that an increase in *desired* investment gets translated into *effective* demand? (The same question, it may be noted, might be asked about any displacement from macroeconomic equilibrium, for instance, the textbook displacement of a short-period equilibrium by a shift in the investment function.)

One possibility is that desired saving increases in line with desired investment, but this assumption in essence requires us to abandon the idea of separate saving and investment schedules. A second possibility that can be dismissed almost as easily is the just-so story fashionable in my youth: we were told to think in terms of cash 'hoards', with 'dishoarding' as the essential mechanism for initiating the disequilibrium transition from one steady state to another. That story won't wash because under contemporary conditions there simply aren't cash hoards of the requisite magnitude – if there ever were.

There *is* a better answer, and interestingly Wicksell, as well as Schumpeter and Keynes, give it in about the same way. Schumpeter is the clearest of the three, in contrast to whom Keynes is practically incoherent, perhaps because he thought the main point of the story so obvious that it did not require elaborate explanation. The main point, in two words, is credit money. The process of expanding investment can get started with no accompanying increase in desired saving if capitalists are assumed to have access to an accommodating banking system which one way or another can create claims on scarce resources out of whole cloth. Here the psychology of financial capital joins the psychology of industrial capital, for unless the financiers share the optimism of industrialists, there is no way, absent those mythical cash hoards, by which investment can increase without a contemporaneous increase in desired saving, as the neo-Keynesian and the Schumpeterian story, with their reliance on price level and profit rate shifts to accommodate capitalists, would have it. Or for that matter, the Wicksellian story. Although ultimately it is the interest rate which adjusts desired investment and saving, Wicksellian disequilibrium is not a Walrasian virtual or hypothetical imbalance of *tâtonnement* with false trading, but an imbalance in real time which is sustained by credit money.

The importance of an accommodating banking system, or passive or endogenous money – these are all approximate equivalents – to the neo-Keynesian system explains why partisans of this view are necessarily hostile to the quantity of money theory of prices (the so-called 'quantity theory of *money*', but this is an obvious misnomer). The dispute is evidently not about the relationship between the quantity of money and the price level, a definitionally true relationship in its customary form, but about causality. In the neo-Keynesian view it is aggregate demand which drives *both* sides of the quantity equation, not MV (the product of money and velocity) which drives PX (the product of prices and quantities).

In an earlier essay on this subject (Marglin 1984a), I suggested that in the neo-Keynesian story capitalists, as a class if not individually, were approximately in the position of the present Aga Khan, who in his student days is reputed to have asked his Harvard economics instructor how the theory of consumer choice worked *without* the budget constraint. V. Bhaskar has persuaded me that the analogy is misleading. The point is not that capitalists, even as a class, face no budget constraint: capitalists must be assumed to repay whatever debts are contracted to finance investment. The point is rather that, as a class, capitalists are able to change what are normally regarded as parameters of the budget constraint. By increasing relative demand, capitalists drive up the prices of the goods they sell relative to costs of production. The consequent increase in profits provides the wherewithal to retire debt as it becomes due.

There is a final assumption which must be introduced in order to make the neo-Keynesian argument that the longer run growth of the

capitalist economy, as well as the share of investment in output and the profit rate, is sensitive to capitalists' animal spirits. This is the assumption of slack resources, specifically, slack labour-force growth must lead to continued capital deepening, which has finite limits both in the real world and in the textbook world of smooth substitution between capital the labour. The simplest, as well as the most realistic, justification of slack labour, is the argument that the capitalist sector of the economy is embedded in a larger entity from which it can, if needed be, draw labour. I refer here to a long run 'reserve army' constituted both by sectors of the national economy (the farm in an earlier day, the kitchen more recently) and by sectors of the international economy (the immigrants who provided labour for 19th- and 20th-century expansion in the United States and for the post-World War II boon in Europe).

Without this assumption, the power of capitalists to shape the history of capitalist economies would be much closer to the power any group has in a standard Arrow–Debreu model to influence relative prices through its preferences or, in a stochastic model, through its expectations. On the assumption of a limited quantity of land suitable for growing tobacco, a shift of smokers' preferences with respect to the pleasures of nicotine and the horrors of lung cancer will affect the equilibrium price of cigarettes. In a stochastic model, the belief that sunspots or any other exogenous variable matters may be sufficient to make the variable matter, even with the assumption of rational expectations (Azariadis 1981; Cass and Shell 1983). At issue in both these cases however is the distribution of a given pie, which stands in sharp contrast to neo-Keynesian theory, where distribution bears on the size of the pie itself.

Observe that the existence of slack resources is much more problematic in the long run than in the short period which was originally the focus of Keynes's analysis. For the short period, although controversy is not lacking even here, mainstream economists will generally accept it is the exception rather than the rule that capacity or manpower constrains output. It is in the long run that the true proportions of the neo-Keynesian departure from orthodoxy reveal themselves.

The Level of Real Wages: Effect or Cause of Accumulation?

In a sense, this version of neo-Keynesian theory, like Keynes's own *General Theory*, proves too much. One cannot ask many of the more interesting questions about how changes in investment demand change the economy. For most of these questions turn out to be about movements along the demand curve rather than changes in its position, which mean that they are necessarily transitory if the initial position was one of equilibrium. For instance, one cannot ask what the effect of lower (or higher) real wages might be because the real wage, determined ultimately by the price level, is a consequence rather than a thermostat. Thus to find a thought experiment which illustrates the neo-Keynesian theory requires some care; it would not have done, for example, to take as the premise of a shift in the investment demand function that Mrs Thatcher and Mr Reagan had abolished collective bargaining. For while this might plausibly lead to the expectation of a higher profit rate, this expectation would translate into a movement *along* the existing investment demand schedule rather than a shift of the schedule. And supposing the economy was at equilibrium in the first place, a movement along the investment demand schedule is not sustainable unless the saving schedule shifts simultaneously.

In fact questions about the effect of changes in real wages on equilibrium come out of a very different approach to capitalism than that embodied in the *General Theory*. There Keynes calls this approach 'classical', even though the economists Keynes apparently had in mind as its exemplars were more neoclassical than classical. Classical may still be the right label since the Marxian strand of the classical theory, as typified by Michal Kalecki, even more than the neoclassical strand, makes the level of real wages a central determinant in the theory of accumulation, a thermostat rather than a thermometer. But if real wages are given exogeneously, whether as a rate or a share, something else has to adjust if both investment and saving propensities are to continue to play a determining role in the accumulation process.

I

There are at least two ways out, besides the possibility of partitioning outcomes by 'regimes' in which one or another consideration – wages, investment, or saving – is a nonbinding constraint. One possibility is to allow each of the three determinants to operate with diminished force; this is the tack followed in Marglin (1984a, b).

Capacity Utility as an Adjustment Mechanism

Another possibility is to allow capacity utilization to enter the discussion more fully both as cause and as effect. This is not entirely unproblematic since it can be argued that the appropriate assumption for the long run (as distinct from Keynes's short period) is that capacity utilization settles down at some 'normal' rate, in other words, that it is not a variable. But it is hard to regard this issue as resolved; both the view that capacity utilization is endogenous and the view that it is exogeneous in the long run can claim some empirical support.

The endogenous view, which Amit Bhaduri and the present author are currently developing, builds indirectly on the work of Kalecki and more directly on the work of Bob Rowthorn (1982). This view entails a significant reformulation of the investment demand function: investment demand is no longer a function of the expected profit rate but rather is a function of two of its constituents, the expected profit *margin* qe and the expected rate of capacity utilization ze. Definitionally $re = qezey$, where y represents the output:capital ratio at full capacity utilization, but with capacity utilization variable, there is no compelling reason why qe and ze should affect investment demand symmetrically, as they do in this formula. If, for example, the expected rate of profit is high because the profit margin is high and the expected rate of capacity utilization is low, the impact on investment demand may be very different from what a high expected rate of profit based on lower profit margins and higher capacity utilization would induce. In the second case there would be relatively little need for investment in order to have sufficient capacity to meet expected demand. Investment demand would be limited to new products, new processes, or the substitution of capital for labour. Thus the formulation of the expected rate of profit from a project as

$$r_i^h = p_1^h r_{i1}(q^e, z^e) + \cdots + p_m^h u_m^h r_{im}(q^e, z^e)$$

is at once more general and more plausible than the formulation in which r_i^h turns on re alone.

In this model, unlike the model in which aggregate investment turns on the expected rate of profit, one *can* ask questions about exogenous shifts in real wages. Profit margins are determined in large part by struggles over real wages, so an increase in the expected wage lowers qe and leads to a reduction in investment demand, that is, to a movement down the investment demand schedule. But the movement may be permanent since there is now another degree of freedom, capacity utilization, to accommodate the change.

A word of caution: the outcome of a change in real wages cannot be predicted solely on the basis of the qualitative structure of the model. Whether higher wages and lower profit margins will increase capacity utilization and accumulation depends on the relative strength of two opposing tendencies. On the one hand, from the capitalists' point of view, a decrease in profit margins makes investment less desirable at a given rate of capacity utilization. On the other hand, a shift in the distribution of income from capital to labour may be expected to increase consumption demand and thus to stimulate capacity utilization. Since investment is by assumption sensitive both to profit margins and to capacity utilization, the relative strength of these two influences becomes crucial. This framework is thus broad enough to accommodate both the 'stagnationist' interpretation of Keynes with its emphasis on high real wages as an engine of accumulation and an 'exhilirationist' interpretation, which emphasizes high profit margins.

Evidently a variety of theories, and an even wider variety of models, is possible within the broad Keynesian vision that aggregate demand matters. The essence of that vision is the role of investor psychology, the strength, in Keynes's words, of 'animal spirits' that 'urge to action

rather than inaction'. Perhaps a paraphrase of Marx puts the main point best: capitalists make the history of capitalist economy, but not in circumstances of their own choosing. Whatever the mechanism of adjustment, whether income distribution or capacity utilization or some combination of the two, capitalists occupy a position of singular privilege in the neo-Keynesian conception, possessing the ability to impress their subjective construction of the future on the present functioning of the economy.

See Also

- ▶ Accumulation of Capital
- ▶ Classical Growth Models
- ▶ Neoclassical Growth Theory

Bibliography

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

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25(3): 380–396.

Cass, D., and K. Shell. 1983. Do sunspots matter? *Journal of Political Economy* 91(2): 193–227.

Harrod, R. 1948. *Towards a dynamic economics*. London: Macmillan.

Kaldor, N. 1956. Alternative theories of distribution. *Review of Economic Studies* 23(2): 83–100.

Kaldor, N. 1957. A model of economic growth. *Economic Journal* 67: 591–624.

Kalecki, M. 1971. *Selected essays on the dynamics of the capitalist economy 1933–1970*. Cambridge: Cambridge University Press.

Keynes, J.M. 1930. *Treatise on money*, The pure theory of money, vol. 1. London: Macmillan.

Keynes, J.M. 1936. *The general theory of employment, interest, and money*. London: Macmillan; New York: Harcourt, Brace & Co.

Keynes, J.M. 1937a. Alternative theories of the rate of interest. *Economic Journal* 47: 241–252.

Keynes, J.M. 1937b. The ex-ante theory of the rate of interest. *Economic Journal* 47: 663–669.

Marglin, S.A. 1984a. *Growth, distribution and prices*. Cambridge, MA: Harvard University Press.

Marglin, S.A. 1984b. Growth, distribution, and inflation: A centennial synthesis. *Cambridge Journal of Economics* 8(2): 115–144.

Marglin, S.A., and A. Bhaduri. 1986. *Distribution, capacity utilization, and growth*. Cambridge, MA: Harvard University.

Myrdal, G. 1939. *Monetary equilibrium*. London: Hodge.

Robinson, J. 1956. *The accumulation of capital*, 2nd ed. London: Macmillan, 1965.

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

Rowthorn, B. 1982. Demand, real wages, and economic growth. *Studi Economici* 18: 3.

Schumpeter, J.A. 1911. *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle*. Trans. R. Opie. Cambridge, MA: Harvard University Press, 1934.

Wicksell, K. 1901. *Lectures on political economy*, General theory, vol. I. Trans. E. Classen. London: Routledge & Kegan Paul, 1934.

Investment Decision Criteria

Jack Hirshleifer

JEL Classifications
E2

Investment is present sacrifice for future benefit. Individuals, firms, and governments all are regularly in the position of deciding whether or not to invest, and how to choose among the options available. An individual might have to decide whether to buy a bond, plant a seed, or undertake a course of training; a firm whether to purchase a machine or construct a building; a government whether or not to erect a dam. Under the heading of investment decision criteria, economists have addressed the problem of how to choose rationally in situations that involve a tradeoff between present and future.

The Economic Theory of Intertemporal Choice

The object of investment is taken to be to optimize one's pattern of consumption over time. The elements needed to determine an individual's investment decision are: (a) his *endowment*, in the form of a given existing income stream over time; (b) his *preference function*, which orders in

desirability all possible time-patterns of consumption; and (c) his *transformation set*, which specifies the possibilities for transforming the original endowment into other time-combinations of consumption.

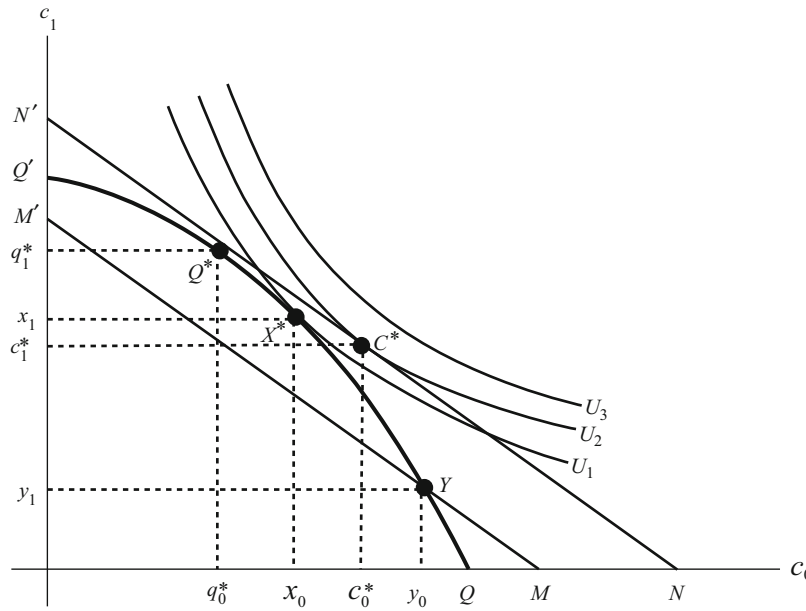
Figure 1 illustrates an artificially simple case of only two periods (say, this year and next) under conditions of certainty. Each point represents a combination of current consumption c_0 and future consumption c_1. The *endowment* combination Y has coordinates (y_0, y_1). *Time-preferences* are portrayed by the indifference curves U_1, U_2, U_3,. . ., each such curve connecting combinations yielding equal satisfaction. The curve QQ' through the endowment position Y pictures the intertemporal *productive opportunities*. By sowing seed, for example, a person can sacrifice current consumption for future consumption – represented in the diagram by a movement from Y along QQ' to the northwest. (There may also be disinvestment opportunities, i.e., the individual might be able to draw upon the future so as to augment current consumption, which would be represented by a movement from Y along QQ' to the south-east.)

For a Robinson Crusoe, the optimum balance of present and future consumption – which in his isolated state must necessarily be identical to his provision for present and future production – occurs at point X^* along QQ'. In the situation pictured he achieves this optimum by *investing* the quantity $y_0 - x_0$ of current consumption claims. For example, having at hand a current corn endowment of y_0, he retains x_0 for current consumption and plants the remainder as seed. Next year he will reap as return from investment the amount $x_1 - y_1$ to augment his endowed availability of future corn.

If markets for trading between present and future income claims exist, however, in contrast with the Robinson Crusoe situation the individual will be able to disconnect the amount he *invests* from the amount he *saves*. These trading opportunities are shown in Fig. 1 by the family of 'market lines' whose general equation is:

$$c_0 + c_1/(1 + r_1) = W_0 \tag{1}$$

Here r_1 is the *interest rate* that discounts one-year future claims c_1 into their equivalent value in terms of c_0 claims. Along each market line the parameter W_0 represents the associated level of *wealth*. Put another way, wealth in Eq. (1) measures the *present worth* of any specified



Investment Decision Criteria, Fig. 1 Investment and saving in a 2-period model

(c_0, c_1) vector – the future-dated element being 'discounted' at the given market interest rate r_1. In the diagram two market lines are shown: MM' through the endowment vector $Y = (y_0, y_1)$ indicates the individual's endowed wealth $W_0^y = y_0 + y_1/(1 + r_1)$, while NN' represents the maximum attainable level of wealth $W_0^* = q_0^* + q_1^*(1 + r_1)$.

If an individual has both productive and market opportunities, his optimizing decision in Fig. 1 can be thought of as taking place in two stages. First he locates his 'productive solution' $Q^* = (q_0^*, q_1^*)$ by moving along QQ' so as to maximize attained wealth at the tangency with market line NN'. Second, he then transacts in the funds market, by lending or borrowing (exchanging current for future claims or vice versa) along NN' to find his 'consumptive solution' $C^* = (C_0^*, C_1^*)$ at the tangency of NN' with indifference curve U_2 in the diagram. Notice that his preferences do not at all affect the productive solution, but only how he chooses to 'finance' the investments made. Specifically, in the diagram here the amount he *invests* $(y_0 - q_0^*)$ exceeds the amount he *saves* $(y_0 - c_0^*)$. By borrowing on the market, in effect he has been able to get others to undertake part of the saving necessary to finance his projected investments.

This disconnection between the individual's productive and consumptive decisions in a regime of perfect markets is known as 'Fisher's Separation Theorem'. The essential implication is that individuals with diverging time-preferences can nevertheless come together and agree upon joint productive investments. Business firms and (to some extent) governments can be regarded as institutions designed for undertaking joint investments whose scale is too large for any single individual. The underlying principle is that those investment choices maximizing wealth value or present worth of the mutual undertaking will also maximize wealth for each and every participant therein.

The Present-Value Rule

The economic theory of intertemporal choice leads immediately to what is known as the *Present-Value Rule* for investment decision. This rule can be expressed in two essentially equivalent forms:

(i) Among the opportunities available, adopt the set of investments that maximizes wealth W_0.
(ii) Adopt any single investment project if and only if its present value V_0 is positive. (Taking into account, of course, any repercussions of that project upon the returns yielded by other members of the adopted investment set.)

As an obvious corollary, if two available projects are mutually exclusive, the one with the larger present value V_0 should be chosen.

Generalizing to the multi-period context, wealth as maximand becomes:

$$W_0 = q_0 + q_1/(1 + r_1) + q_2/[(1 + r_2)(1 + r_1)] + \cdots + q_T/[(1 + r_T)\cdots(1 + r_2)(1 + r_1)] \quad (2)$$

Here the q_t are the coordinates of points along the T + 1-dimensional productive opportunity surface

$$\varphi(q_0, q_1, \ldots, q_T) = 0,$$

a generalization of curve QQ' in Fig. 1. T is the 'economic horizon', which may be infinite. And the r_t represent the successive short-term interest rates, each of which discounts prospective payments at any date into its wealth-equivalent at the next preceding date.

For a single project in the multi-date context, present value is defined as:

$$V_0 = z_0 + z_1/(1 + r_1) + z_2/[(1 + r_2)(1 + r_1)] + \cdots + z_T/[(1 + r_T)\cdots(1 + r_2)(1 + r_1)] \quad (3)$$

Here the z_t are the dated payments or 'cash flows' associated incrementally with the project considered. Normally the z_1 elements for earlier dates would include some with negative signs – or else the project could not be described as an investment – while those for later dates would have predominantly positive signs. In the special case where $r_1 = r_2 = \ldots = r_T = r$ – that is, where

interest rates are expected to remain constant at the level r over time – the Present-Value formulas reduce to the more familiar forms:

$$W_0 = q_0 + q_1/(1+r) + q_2/(1+r)^2 + \cdots + q_T/(1+r)^T \quad (2')$$

$$V_0 = z_0 + z_1/(1+r) + z_2/(1+r)^2 + \cdots + z_T/(1+r)^T \quad (3')$$

The Present-Value solutions can also be formally generalized to allow for continuous rather than discrete time. As an illustrative simplified example, consider a project whose scale of current input or investment sacrifice i_0 is fixed while the output date is subject to choice (e.g., when to cut a growing tree). In Fig. 2, horizontal distances represent time t and vertical distances value V_t at each date. Present Value V_0 is indicated by height along the vertical axis. The curve GG' represents productive growth of the asset – in the case of a tree, market value of the standing timber at any date. The 'discount curves' $D, D', D'', \ldots$, are analogous to the 'market lines' of Fig. 1. Each such curve represents the growth of a specific sum of present dollars by continuous compounding at a constant market rate of interest r, or alternatively the Present Value of any future payment continuously discounted at r. The optimal investment period $t = t^*$ is then the one that maximizes Present Value V_0, subject to the constraint on the available V_t described by the curve GG', in the equation:

$$V_0 = -i_0 + V_t e^{-rt} \quad (4)$$

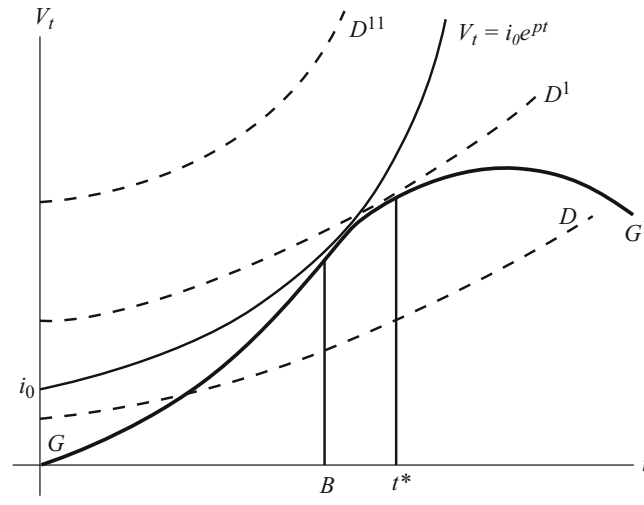
Geometrically, t^* is determined by the tangency of GG' with the highest discount curve (constant-wealth curve) attainable. The solution condition is then:

$$V_t'/V_t = r \quad (5)$$

Other Investment Criteria

Certain investment criteria employed in business practice are definitely erroneous. One such is rapidity of 'payout' (the date when cash inflows first balance initial outlays), a formula that obviously fails to allow properly for time-discount. Controversy among theorists has centred upon a more interesting concept known variously as the 'internal rate' or the 'rate of return'. The internal rate for a project (or set of projects) is defined as ρ in the discrete discounting equation:

$$0 = z_0 + z_1/(1+\rho) + z_2/(1+\rho)^2 + \cdots + z_T/(1+\rho)^T \quad (6)$$



Investment Decision Criteria, Fig. 2 Optimal during of investment

As before the z_t here are the successive terms, positive or negative, of the paymentsreceipts sequence associated incrementally with a particular project. In the special 'deepening' case illustrated in Fig. 2, the corresponding concept under continuous compounding is defined implicitly in:

$$0 = -i_0 + V_t e^{-\rho t} \tag{7}$$

where once again the V_t at any date is described by the productive opportunity curve GG'. Under these conditions ρ represents an average compounded rate of growth.

There has been some confusion between two quite different investment decision *rules* that both employ the internal-rate measure ρ: (i) choose projects so as to maximize ρ, versus (ii) adopt projects incrementally so long as $\rho > r$.

Maximum ρ Rule

If the internal rate ρ is interpreted as the average rate of growth, it may seem plausible that the investor should maximize ρ rather than wealth W_0. (Of course, maximizing a growth *rate* would scarcely make sense unless the initial outlay or scale of investment were held constant, which would not in general hold true.) The solution of (7) that maximizes ρ is shown in Fig. 2 as $t = B$, notably earlier than the Present-Value solution $t = t^*$.

In favour of B over t^* it has been argued that, if the growth opportunity were to be replicated in perpetuity, returns from choosing the earlier 'rotation period' B must ultimately dominate those associated with cutting on each cycle at t^*. That is certainly true. However, if the decision problem concerns infinite rotation rather than a one-time cutting, for a valid comparison the relevant Present-Value measure would have to be a generalized one that allows for the associated infinite sequence of discounted returns. It can be shown that this generalized Present-Value does coincide with B if the growth opportunity can be reproduced on an ever-broadening scale (e.g. on new land) – but only as funds are freed by cutting the tree or trees. This turns out to be an impossible or uninteresting case, because it implies that the productive opportunity must be of *infinite* market value if the maximized ρ exceeds the market interest rate r (and of zero value otherwise). In contrast, if the opportunity is a unique one which cannot be reproduced after cutting, as pictured in Fig. 2, the simple $t = t^*$ solution remains correct. Another solution, $t = F$, found by the German forester Faustmann, is appropriate when the opportunity can be reproduced over time by cutting and replanting but cannot be broadened in scale. F would be found by maximizing the Present Value V_0 of an infinite sequence of rotations, each being a constant-scale replication of the original opportunity. Like all the correct solutions, it is equivalent to maximizing the present worth of the opportunity under the stated assumptions. (F is not shown in Fig. 2 but would lie between B and t^*.)

ρ vs. r Comparison Rule

The Comparison Rule says to adopt any project whose internal rate ρ exceeds the market rate of interest r. This rule remains popular in business practice, in part because it offers a convenient division of labour: calculation of the ρ's on individual projects might be delegated to subordinates, while top decision-makers choose the cutoff rate r that corresponds to the relevant market interest rate faced by the firm. Unfortunately, however convenient such a decision of labour may be, once again this is not in general a correct method of project selection.

The difficulty with the Comparison Rule first came to be appreciated when it was discovered that a sequence of positive and negative cash flows could have more than one ρ serving as solution of Eq. (6) above. A project represented by the annual payments sequence $-1, 5, -6$, for example, has two solutions: $\rho = 1$ and $\rho = 2$. (It can be shown that a project with $T + 1$ dated elements may have as many as T solutions.) This of course destroys the idea that the internal rate can generally be identified with a growth rate; an outlay of one dollar cannot be said to grow at both 100% and 200%. Various answers have been offered to the puzzle of which ρ to use in such cases. But the difficulty is

immediately explained and resolved if we think instead in terms of Present Value. It turns out that the sequence $-1, 5, -6$ has positive V_0 (and is therefore worth adopting) for any constant market interest rate r *between* 100% and 200%, but at other values of r has negative Present Value (and should not be adopted). Perhaps even more illuminating is the project described by cash flows $-1, 3, -2\frac{1}{2}$. This sequence has no real solution for ρ in Eq. (6), the reason in Present-Value terms being that V_0 investment opportunity. After all, there is no justification for postulating (as is implicitly done by the Comparison Rule) that the anticipated sequence of market interest rates $r_1, r_2, \ldots, r_T$ must be constant over time (always equal to a common r). It turns out that the cash-flow pattern $-1, 3, -2\frac{1}{2}$ has positive Present Value (i.e., the project would be worth adopting) for many possible non-constant interest-rate sequences – for example, $r_1 = 100\%$ and $r_2 = 200\%$.

Summing up, therefore, the Present-Value Rule for investment decision – corresponding as it does to the principle of maximizing wealth within the opportunities available – is correct itself and also serves to define the range of validity of all the other rules considered.

Generalizations and Extensions

The preceding analysis needs to be extended in at least two important ways, so as to allow for: (1) uncertainty, and (2) imperfect and incomplete markets.

Uncertainty

Investment choices, involving as they do present sacrifice for future benefit, are peculiarly sensitive to uncertainty. However, so long as we can continue to assume a regime of complete and perfect markets, the Present-Value Rule is robust enough to retain validity even in a world of uncertainty. For, the proximate goal of any individual (or group of individuals organized in a firm or other joint enterprise) will still be to undertake productive activities so as to maximize wealth. Having achieved that goal, each and every individual investor will be in a position to distribute his attained wealth as desired over all possible dated contingencies in accordance with his time-preferences, degree of risk-aversion, and probability beliefs.

Economists use two main models for the analysis of uncertainty – state-preference and mean-versus-variability analysis. Since the latter, under certain assumptions, can be regarded as a special case of the former, for our purposes attention can be limited to the state-preference model. If markets for state-claims are complete and perfect, any pattern of varying returns over states of the world at a given date has a *certainty-equivalent* in value terms as of that date. In Eqs. (3) and (3′), the z_t for any project can now be interpreted as certainty-equivalents (rather than as simple cash flows) defined by:

$$z_t = P_{t1}z_{t1} + P_{t2}z_{t2} + \cdots + P_{tS}z_{tS} \qquad (8)$$

Here z_{ts} represents the cash flow at date t contingent upon state of the world s obtaining – there being S distinguishable such states – while P_{ts} is the price at which a unit claim to income in state s at date t can be converted into (traded for) current certainty income.

Incomplete or Imperfect Markets

Markets are said to be *incomplete* if some objects of choice are non-tradeable. For example, futures markets for some commodities at far-distant dates do not exist, nor is it possible to trade in claims contingent upon each and every conceivable future uncertain event. Markets are said to be *imperfect* if there are costs of trading – for example, brokerage fees, transaction taxes, or expenses in locating exchange partners. Any real-world regime of markets will necessarily be both incomplete and imperfect, but for some purposes the assumption of complete and perfect markets may be a usable idealization. Unfortunately, once we depart from this idealization the problem of investment decision criteria becomes very difficult. The reason is that the Separation Theorem fails. Only under complete and perfect markets is the concept of wealth or Present-Value unambiguously defined, so that the choice of productive investments can be entirely disconnected from

individuals' personal time-preferences, risk-preferences, beliefs etc. Failure of the Separation Theorem particularly subverts the ability of investors to join together in undertaking large projects or groups of projects.

However, two different lines of analytical approach have yielded results of interest. (i) A number of techniques have been devised for locating 'utility-free' or 'efficient' investment choices. In general such techniques cannot determine an optimal project set, but they can serve to filter out options whose payoff patterns over dates and/or states are dominated by other available projects or project combinations. (ii) While investors' personal circumstances may diverge in innumerable ways, there should be some tendency for those similarly situated to group together. Thus, a firm whose investment opportunities yield far-future payoffs should tend to be owned by a 'clientele' consisting of individuals with moderate time-preferences, willing to forego current dividends in the hope of large long-term gain. It follows that unanimity as to the investment choices to be made may after all govern *within* the firm, for example as to the discount rate to employ in calculating Present Value, even in the absence of perfect and complete markets.

See Also

- ► Internal Rate of Return
- ► Present Value

Bibliography

The modern theory of investment and intertemporal choice was set down in classic form by Irving Fisher as part of his great works on interest (1907, chapters 8–9; 1930, chapters 10–13). The seminal works on uncertainty theory include Arrow (1953) for the state-preference approach and Markowitz (1959) and Sharpe (1964) for the mean-versus-variability model. Choice over time and choice under conditions of uncertainty are integrated in the treatise by Hirshleifer (1970) that builds upon these foundations. All these topics have been followed up in an enormous literature, of which only a few illustrative instances can be cited here: on investment decision formulas, Samuelson (1976); on utility-free or dominant choices, Pye (1966), Hanoch and Levy (1969), and DeAngelo (1981). A survey of investment decision criteria used in current business practice appears in Schall, Sundem, and Geijsbeek (1978).

Arrow, K.J. 1953. The role of securities in the optimal allocation of risk-bearing. Reprinted in K.J. Arrow, *Essays in the theory of risk-bearing*. Chicago: Markham, 1971.

DeAngelo, H. 1981. Competition and unanimity. *American Economic Review* 7 (1): 18–27.

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

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Hanoch, G., and H. Levy. 1969. Efficiency analysis of choices involving risk. *Review of Economic Studies* 36 (3): 335–346.

Hirshleifer, J. 1970. *Investment, interest, and capital*. Englewood Cliffs: Prentice-Hall.

Markowitz, H.M. 1959. *Portfolio selection*. New York: Wiley.

Pye, G. 1966. Present values for imperfect capital markets. *Journal of Business* 39 (January): 45–51.

Samuelson, P.A. 1976. Economics of forestry in an evolving society. *Economic Inquiry* 14 (4): 466–492.

Schall, L.D., G.L. Sundem, and W.R. Geijsbeek. 1978. Survey and analysis of capital-budgeting methods. *Journal of Finance* 33 (1): 281–287.

Sharpe, W.F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19 (3): 425–442.

Investment Planning

Joseph Halevi

The theories discussed here consist of two complementary formulations originating in India and in the United Kingdom in the 1950s and 1960s. Both deal with investment planning when development starts with virtually no capital goods industry. Thus they represent an expansion of the model of the Soviet economist Feld'man, since in the latter the economy did possess an investment sector albeit in a limited dimension (Feld'man 1928a, b).

The first approach, due to Dobb (1954, 1960) and to Sen (1960), deals with the choice of techniques and the sectoral distribution of investment and labour. The second approach, elaborated by a number of Indian scholars – Raj and Sen (1961),

Naqvi (1963) – is more concerned with the sectoral allocation of investment goods under conditions of stagnant export earnings. The definition of sector is the same as in the Marx-based Feld'man model with the difference that the capital goods sector itself is divided into two branches. One branch consists of an intermediate sector producing equipment usable only in the consumption goods sector. The second branch is formed by *machine tools* which can reproduce themselves as well as be installed in the intermediate sector.

The emphasis on this kind of structural relations is aimed at providing analytical support to the view that sectoral investment planning by the State is a necessary, although not sufficient, condition for the emancipation from backwardness.

The starting point of both approaches is the historical consideration that colonialism has destroyed the traditional home industries, thereby making expansion dependent on the exports of primary products having low demand elasticities (Raj and Sen 1961). It is this particular condition which justifies investment priority in the capital goods industry for a growth strategy oriented toward the home market (Dobb 1967). Industrialization would then imply the creation of capital goods well in advance of any market demand for them, a process called by Dobb *the Accelerator in Reverse*.

Developing economies face the task of investing in a manner largely independent from the preexisting material structure. In this context, indivisibilities of capital equipment – which 'are likely to be significantly large (relatively to the scale of the economy) at early stages of development' (Dobb 1960, pp. 11–12) – may make the expansion of a certain branch unprofitable although its growth can be of crucial importance for the formation of other industries. State planning of the sectoral allocation of investment performs the role of securing overtime the construction of complementary industries.

It must be noticed that some of the views put forward by Dobb and the Indian economists were part of the intellectual climate of the period. In the mid-1950s Prebisch started the debate over the terms of trade between industrialized and underdeveloped countries, arguing the long-term nature of the latter's unfavourable position. Politically, the first meeting of the non-aligned nations, held in the Indonesian city of Bandung in 1955, asserted the necessity to embark on a road privileging the domestic market. Institutionally, sectoral planning by the State seemed to have gained a firm hold also in a non-socialist country as important as India. Practically, the experience of the People's Republic of China suggested that a developing country could reduce the dependency on foreign exchange by building a machine tools industry (Raj 1967).

Given this cultural and political framework, Dobb's pioneering work has a special place in the theories of planned development. It singled out the fact that the domestic economy of underdeveloped countries does not generate a surplus of wage goods large enough to allow a more or less smooth process of growth. Indeed, with most of the work force employed in subsistence activities, it would be impossible to set in motion the Accelerator in Reverse unless the bottleneck of a limited surplus is widened. The technical form of investment must therefore reflect this initial constraint. In setting forth the answer to the question of the choice of techniques, Dobb challenged the view that 'since a scarcity of capital relative to labour is a usual characteristic of underdeveloped economies, capital investment needs there to take the form of projects of "low capital intensity"' (Dobb [1954] 1955, p. 139).

The gist of his and Sen's argument (Sen 1960) can be presented as follows:

Consider an economy where fixed capital in the capital goods industry is so small that machines can be thought of as being produced by labour alone. Thus, employment in the capital goods sector multiplied by the productivity of labour – denoted by x – gives the total output of equipment. But employment in the capital goods sector is limited by the surplus produced in the wage goods sector. If 20 people work in the wage goods sector, where the productivity of labour (z) is 20 units per person and the real wage rate (w) is uniform throughout the economy and fixed at 10 units, then 20 people can be put to work in the capital goods sector. The crucial ratio is given

by $(z - w)/w$, where $z - w$ is the surplus per unit of labour in the wage goods sector. If the bottleneck in the production of wage goods has to be widened without lowering the real wage, all newly produced machines should be installed in the wage goods sector. On the assumption that these do not depreciate and that each machine employs one worker, total output of capital goods will be equal to the increment in employment in the wage goods sector. The growth rate of the economy is therefore equal to the growth rate of employment in this sector. Given the above mentioned allocation policy, the growth rate is nothing but the productivity of labour in the capital goods sector multiplied by the ratio of the surplus to the wage rate. Hence:

$$g = x(s/w); \quad \text{where} \quad s = z - w. \tag{1}$$

Assuming no production lags, maximization of (1) yields:

$$-(\mathrm{d}x/x) = (\mathrm{d}z/z)(z/s). \tag{2}$$

According to Eq. (2), the growth rate would be maximized by using more costly methods of production in the capital goods industry, lowering the productivity of labour in this sector. At the same time, the delivery of improved and more expensive equipment would *ipso facto* raise labour productivity in the wage goods industry. With a positive wage rate – implying a *z/s* ratio greater than unity – this gain need not be as large as the loss of productivity in the capital goods industry. It is the asymmetrical change in the sectoral productivities of labour which leads to an overall increase in capital intensity.

The results do not change if unassisted labour builds machine tools for the intermediate investment sector. In this case the gains in the intermediate sector multiplied by *z/s*, should equal the losses in the machine tools industry.

With a construction based on a number of simplifying assumptions, Dobb and Sen provided the rationale for raising the capital intensity of production under conditions of abundant labour supply. Yet the assumptions turned out to be restrictive not so much in relation to traditional theory, but in relation to the scope and objective of the exercise.

Analytically the model does not succeed in giving a criterion for the choice of techniques when the economy embarks on a path of self expansion of the machine tools sector. The only possible observation is that this sector's productivity does not depend on any other branch of the economy, thus there is no constraint on the degree of capital intensity (Johansen and Ghosh, in Dobb 1960). Dobb's and Sen's results depend very much on the assumptions of no production lags and of immortal machines. In macroeconomic terms, an increase in capital intensity generates a higher growth rate only if the share of investment in national income is raised more than proportionately, which may not be immediately feasible. In the interim period the economy will experience a lower growth rate and a lower share of consumption (Kalecki 1972a). In turn, the notion of immortal machines becomes untenable whenever Dobb analyses the possibility of drafting the whole of the labour force in the two investment industries for the purpose of building the machine tools sector. If wear and tear is taken into account, as soon as no equipment flows to the wage goods sector its capital stock will shrink and so will the output of consumables. The wage rate will cease to be a parameter, becoming instead a variable conditioned by the proportions in which labour and machines are distributed. Hence, wear and tear and the socially minimum wage rate show the limit of the percentage in which machine tools can be reinvested in their own sector. This is a major structural and social aspect of any process of accelerated accumulation (Lowe 1976; Halevi 1981).

Dobb's contribution will remain a classic in the field because it introduced a novel perspective on the reasons for, and the modalities of, socialist-oriented development for the ex colonial countries. The fact that this approach is no longer followed can only in part be attributed to the limitations outlined above. Perhaps, in addition to the ever present ideological factor, one explanation lies in the change of the historical framework. There are, by now, significant instances in which a process of fast accumulation has taken

place hand in hand with the persistence of phenomena such as landlessness and urban poverty. In countries like Brazil, Mexico and India, these are the problems that must be reflected in any planning strategy. The issue is not so much that of building a capital goods sector from scratch, but to conceptualize the economic and political nature of the phenomena (Kalecki 1972b, c; Taylor and Bacha 1976).

The second approach, coming mainly from India, is a substantial improvement on the Mahalanobis variant of the Feld'man model (Mahalanobis 1955). It uses the same hypothesis of two capital goods industries to discuss the sectoral allocation of machinery imported through a fixed sum of foreign earnings *F*. Raj and Sen (1961) assumed negligible amount of equipment in the intermediate investment sector *I* and in the machine tools industry *M* Furthermore, machine tools are used also for the extraction of raw materials *R*. The planners can freely choose the initial share of consumption over national income, production coefficients are given. In this context, if *F* is used to import *I* goods for the production of consumption goods *C*, the output of *C* goods will rise but its absolute increase will tend to nought because raw material requirements will also rise. A constant increment in *C* goods production can be obtained when *F* is used to import *M* goods for the production of *I* goods and for the extraction of *R*. In this case raw materials set a limit to the expansion of the *I* sector output. Finally, the output of consumption goods will grow at a constant absolute rate if *M* goods are imported in order to produce machine tools to be installed exclusively in the *I* and *R* sectors.

The original Raj–Sen paper did not discuss the proportions in which machine tools are reinvested in the *M* sector itself. In the literature that followed, the point was raised by Naqvi (1963) and later by Cooper (1983). Naqvi noted that reinvestment in the *M* goods sector would allow for a proportionate growth in *C* goods also in the presence of a limited amount of import earnings. Moreover he observed that central control of the *M* goods sector can be used to limit the creation of a luxury goods industry catering for the well to do. Cooper, on his part, argued that planners can more effectively influence the share of consumption by selecting the ratio in which *M* goods are to be reploughed in their own sector. This is because the share of consumption over national income cannot be freely determined by planners, since it is fixed by the initial distribution of equipment. Planning models based on sectoral relations and on the principle of the Accelerator in Reverse, showed a greater longevity than choice of techniques models. The assumption of given production coefficients did not prevent the analysis of alternative growth paths and the introduction of limiting conditions such as minimum wage rate and stagnant export earnings (Das 1974). The capital goods-consumption goods model has been used also as a framework for the application of optimal control theory in development planning (Stoleru 1965), as well as for the analysis of unused capacity caused by a slow growing agricultural output (Patnaik 1972; Raj 1975).

Contributions to investment planning using analytically a Marxian sectoral approach have come mostly from Great Britain and from India. The Soviet mathematical economists seem to be more inclined toward generic multisectoral optimisation models. This may reflect a belief that a purely capital goods-consumption goods approach ceases to be relevant when a socialist economy possesses a developed industrial structure. Yet, as it emerges from reading the works of some Soviet economists of the mathematical school, generic multisector models cannot give a stylized picture of growth paths (Dadayan 1981). Indeed, in the Western literature on growth, the crucial issue of the transition between two growth rates – a process called Traverse – is dealt with an analytical apparatus closer to Marx's sectoral characterization of the economy (Hicks 1965; Lowe 1976).

See Also

- ▶ Cost–Benefit Analysis
- ▶ Development Planning
- ▶ Planned Economy
- ▶ Project Evaluation

Bibliography

Cooper, C. 1983. Extensions of the Raj-Sen model of economic growth. *Oxford Economic Papers* 35(2): 170–185.

Dadayan, V. 1981. *Macroeconomic models*. Moscow: Progress.

Das, R.K. 1974. *Optimal investment planning*. Rotterdam: Rotterdam University Press.

Dobb, M. 1954. A note on the so-called degree of capital-intensity of investment in under-developed countries. In *On economic theory and socialism, collected papers*, ed. M. Dobb. London: Routledge & Kegan Paul, 1955.

Dobb, M. 1960. *An essay on economic growth and planning*. London: Routledge & Kegan Paul.

Dobb, M. 1967. The question of 'investment-priority' for heavy industry. In *Papers on capitalism, development and planning*, ed. M. Dobb. New York: International Publishers.

Feld'man, G. 1928a. K teorii tempov narodnogo dokhoda. (On the theory of growth rates of the national income.) *Planovoe khoziaistvo*, November.

Feld'man, G. 1928b. K teorii tempov narodnogo dokhoda. (On the theory of growth rates of the national income.) *Planovoe khoziaistvo*, December.

Halevi, J. 1981. The composition of investment under conditions of non uniform changes. *Banca Nationale del Lavoro Quarterly Review* 34(137): 213–232.

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

Johansen, L., and A. Ghosh. 1960. Appendix: notes to chapters III and IV. In *An essay on economic growth and planning*, ed. M. Dobb. London: Routledge & Kegan Paul.

Kalecki, M. 1972a. The problem of choice of the capital-output ratio under conditions of an unlimited supply of labour. In *Selected essays on the economic growth of the socialist and the mixed economy*, ed. M. Kalecki. Cambridge: Cambridge University Press.

Kalecki, M. 1972b. Problems of financing economic development in a mixed economy. In *Selected essays on the economic growth of the socialist and the mixed economy*, ed. M. Kalecki. Cambridge: Cambridge University Press.

Kalecki, M. 1972c. Social and economic aspects of 'intermediate regimes'. In *Selected essays on the economic growth of the socialist and the mixed economy*, ed. M. Kalecki. Cambridge: Cambridge University Press.

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

Mahalanobis, P.C. 1955. The approach of operational research to planning in India. *Sankhya* 16: 3–131.

Naqvi, K.A. 1963. Machine-tools and machines: A physical interpretation of the marginal rate of saving. *Indian Economic Review* 6(3): 19–28.

Patnaik, P. 1972. Disproportionality crisis and cyclical growth. A theoretical note. *Economic and Political Weekly* 7(annual number): 329–336.

Raj, K.N. 1967. Role of the 'machine-tools sector' in economic growth. In *Socialism, capitalism and economic growth, essays presented to Maurice Dobb*, ed. C. Feinstein. Cambridge: Cambridge University Press.

Raj, K.N. 1975. Linkages in industrialization and development strategy; some basic issues. *Journal of Development Planning* 8: 105–119.

Raj, K.N., and A.K. Sen. 1961. Alternative patterns of growth under conditions of stagnant export earnings. *Oxford Economic Papers* 13: 43–52.

Sen, A.K. 1960. *Choice of techniques*. Oxford: Basil Blackwell.

Stoleru, L. 1965. An optimal policy for economic growth. *Econometrica* 33: 321–348.

Taylor, L., and E. Bacha. 1976. The unequalizing spiral: A first growth model for Belindia. *Quarterly Journal of Economics* 90(2): 197–218.

I

Invisible Hand

Mark Blaug

Abstract

Adam Smith employed the term 'invisible hand' twice in his published writings, and a considerable secondary literature has explored the multiple meanings he intended to convey by the use of this metaphor. I argue that, whatever he did mean, he certainly did not mean that competition or the market mechanism promoted efficiency: instead it promoted the growth of income, even for the poor.

Keywords

Austrian economics; Capital accumulation; Classical competition; Competition; Competition as rivalry; Cournot, A.; Division of labour; Ferguson, A.; First fundamental theorem of welfare economics; Hume, David; Invisible hand; Mandeville, B.; Mercantilism; Perfect competition vs free competition; Productive and unproductive labour; Scottish Enlightenment; Smith, A. on invisible hand; Spontaneous order; Srewart, D.; Turgot, A.; Welfare economics

JEL Classifications
B1; B3

The 'invisible hand' was a metaphor used by Adam Smith to describe 'the principle by which a beneficent social order emerged as the unintended consequence of individual human action'. This is Vaughn's succinct summary of Smith's intentions in employing the metaphor (1987, p. 997). More recently, Grampp (2000) has reviewed nine different interpretations of the famous metaphor, concluding that the three references to the invisible hand in Smith's works are not expressions of the same concept, an opinion share by many other commentators.

Smith referred to the 'invisible hand' twice in his published writings (there is a third reference in his unpublished 'Essay on Philosophical Subjects'), and he did so at greatest length in Book IV, Chapter 2, of the *Wealth of Nations*. It is easier to say what he did *not* mean by the invocation to the 'invisible hand' than to spell out precisely what he did mean. What he definitely did not mean is the so-called first fundamental theorem of modern textbook welfare economics, although that reading has been frequently ascribed to him (for example, Arrow and Hahn 1971, p. 1; Mas-Colell et al. 1995, pp. 308, 327, 524, 545, 549). The first fundamental theorem states that, subject to certain exceptions such as externalities, economies of scale, public goods and imperfect information, every competitive equilibrium is Pareto-optimally efficient. It is indeed possible to find statements in Chapter 2 of the *Wealth of Nations* on mercantilism that appear to endorse something like the first fundamental theorem. Capitalists have a preference for domestic over foreign investment for reasons of security, Smith asserts, but the result of the free movement of capital is nevertheless of benefit to society as a whole:

> As every individual, therefore, endeavours as much as he can both to employ his capital in support of domestic industry, and so to direct that industry that its produce may be of the greatest value; every individual necessarily labours to render the annual revenue of the society as great as he can. He generally, indeed neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestic to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this as in many other cases, led by an invisible hand to promote an end which was no part of his intention. (Smith 1776, pp. 455–6)

The natural interpretation of this passage is, at least for domestic industry, that total product is maximized by free competition. This is almost the first fundamental theorem – but not quite.

First, a presumption of maximization is not a mathematical theorem and, secondly and more significantly, free competition or free unrestricted entry into industries is a far cry from *perfect* competition without which the notion of the price-taking behaviour of numerous, small competitors, adjusting only the quantities they buy or sell, falls to the ground. Cournot invented the concept of perfect competition de *novo* in 1838 and, since the proof of the first fundamental theorem absolutely requires the concept of perfect competition, the idea that Adam Smith somehow stated a primitive version of the first theorem must be wrong; it is in fact a historical travesty.

Adam Smith clearly believed in competition, or rather 'the simple system of natural liberty', but his idea of competition was a behavioural one, not defined by the number of firms in the market as in Cournot. Competition, for Smith as for all the classical economists, implied rivalry by price and non-price means, rivalry among consumers bidding for a limited supply and rivalry among producers to dispose of that supply on the most advantageous terms. In other words, he had what I have called a 'process conception of competition', nowadays associated with Austrian economics, in contrast to the orthodox conception of economics, in which all the emphasis is directed to the nature of the final equilibrium, regardless of how that final equilibrium is attained (Blaug 1997, p. 678; see also Coase 1997, p. 318; Kirzner 2000).

Although the first theorem cannot be found in the *Wealth of Nations*, what can be found is the notion that competition has desirable properties, namely, that it promotes the rate of growth of

national income or what he labelled 'the wealth of nations', which results in the material improvement of the standard of living even of the poorest members of society. This idea is not only the mainspring of the famous opening chapter of the book on 'The Division of Labour' in the pin factory, but it accounts for the emphasis on capital accumulation and the crucial distinction for Smith's theory of economic growth between 'productive and unproductive' labour in Book II, not to mention the content of the whole of Book III with its revealing title 'The Different Progress of Opulence in Different Nations', which translated into modern jargon reads 'On Differences in the Growth Rate of Different Countries'. Much of Book III is devoted to persuading the reader that there had been material progress since Elizabeth I, a thesis which surprisingly was frequently denied at the time. In short, what was good about what he called 'the commercial society' was that it grew rapidly, not that it was efficient, a term and indeed a concept that never appears in the *Wealth of Nations*.

Smith's references to an 'invisible hand' in the *Wealth of Nations* have attracted an enormous secondary literature (for example, Hayek 1973, ch. 2; Vaughn 1987; Persky 1989; Grampp 2000; Rothschild 2001, pp. 116 ff. Streissler 2003, Minowitz 2004; Vivenza 2005), no doubt because they express three closely connected but separable ideas: (*a*) the *private* actions of individuals can have unforeseen and unintended *social* consequences; (*b*) these private self-interested actions and unintended social consequences may be harmonious in mutually promoting the interests of all members of society; and (*c*) there is an order in these harmonious outcomes as if private self-interested actions were centrally coordinated to produce a coherent overall pattern. This is a profound assembly of ideas that captures the doctrine of 'spontaneous order' employed by many thinkers of the Scottish Enlightenment to explain the emergence of such social institutions as language, the law, private property, the monetary system and even the market mechanism itself, not by central design or collective regulation but by individual action undertaken for quite different reasons. It arises most clearly in Adam Ferguson's *Essay on the History of Civil Society* (1767), published a decade before the *Wealth of Nations,* and even earlier in Hume's *Treatise of Human Nature* (1740). But important as the idea of a 'spontaneous order' may have been to Ferguson and Hume, as well as to Mandeville, Turgot and Dugald Stewart, it was not actually in the forefront of Adam Smith's thinking and, in any case, he never characterized the price system or even free competition as an 'invisible hand'. This is a modern reading of Smith under the influence of Walras and Pareto as translated by Arrow and Debreu.

It was only in the last quarter of the 19th century (as a result of German critics of Smith) that the phrase 'invisible hand', which after all occurs only once in the *Wealth of Nations*, was elevated to a proposition of profound significance. Rothschild deals expertly with the subject and concludes that 'the image of the invisible hand is best interpreted as a mild ironic joke' (2001, p. 116). This may be going a little too far in the opposite direction to the now prevailing interpretation, but there is no doubt that Smith himself did not attach great importance to the idea of an invisible agency channelling the behaviour of self-interested individuals and instead regarded the metaphor of the invisible hand as a sardonic, if not ironic, comment on the self-deception of all of us, including moral philosophers.

Support for this view of his intentions is found in the one reference to the 'invisible hand' in The *Theory of Moral Sentiments,* a reference that is frequently ignored in the exegetical literature on Smith. In that passage in the *Theory of Moral Sentiments* (1759, pp. 184–5) Smith argues that mankind has progressed in the face of pronounced and persistent inequalities and that the rich, despite their natural selfishness, end up unintentionally sharing their wealth with the poor, who for their part end up no worse than the rich themselves. Both Grampp and Minowitz, alone among all the Smithian commentators, object to this conclusion as too Panglossian. Be that as it may, this passage soon dispels the belief that Smith meant one thing and one thing only by the metaphor of 'the invisible hand'.

The notion of a spontaneous order in the sense of a self-regulating system accounting for the

existence of economic institution went underground after the Scottish Enlightenment, and references to an 'invisible hand' are rarely encountered in any of the classical economists, although the idea that economics studies an underlying invisible reality beneath the surface appearance of a free market economy continued to dominate the thinking of Ricardo, J.S. Mill and particularly Karl Marx.

See Also

- ▶ Competition
- ▶ Cournot, Antoine Augustin (1801–1877)
- ▶ Division of Labour
- ▶ Efficient Allocation
- ▶ Hayek, Friedrich August von (1899–1992)
- ▶ Hume, David (1711–1776)
- ▶ Mandeville, Bernard (1670–1733)
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Pareto, Vilfredo (1848–1923)
- ▶ Spontaneous Order

Bibliography

Arrow, K., and F. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Blaug, M. 1997. *Economic theory in retrospect*. 5th ed. Cambridge: Cambridge University Press.

Coase, A. 1997. The wealth of nations. *Economic Inquiry* 15: 309–325.

Grampp, W. 2000. What did Adam Smith mean by the invisible hand? *Journal of Political Economy* 108: 441–465.

Hayek, F. 1973. *Law, legislation and liberty. volume 1: Rules and order*. London: Routledge & Kegan Paul.

Kirzner, I. 2000. Competition and the market process: Some doctrinal milestones. In *The process of competition*, ed. J. Kraft. Cheltenham: Edward Elgar.

Mas-Colell, A., M. Whinston, and R. Green. 1995. *Microeconomic theory*. New York: Oxford University Press.

Minowitz, P. 2004. Adam Smith's invisible hands. *Econ Journal Watch* 1: 381–412.

Persky, J. 1989. Retrospectives: Adam Smith's invisible hands. *Journal of Economic Perspectives* 3 (4): 195–201.

Rothschild, E. 2001. *Economic sentiments: Adam Smith, condorcet and the enlightenment*. Cambridge, MA: Harvard University Press.

Smith, A. 1759. *The theory of moral sentiments*, ed. D. Raphael and A. Macfie. Oxford: Clarendon Press, 1976.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. R. Campbell, A. Skinner and W. Todd. Oxford: Clarendon Press, 1976.

Streissler, E. 2003. Adam Smith's ultimate invisible hand: Content and rhetoric. *History of Economic Thought Newsletter* 8: 3–15.

Vaughn, K. 1987. Invisible hand. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman. Basingstoke: Palgrave.

Vivenza, G. 2005. The agent, the actor, and the spectator. *History of Economic Ideas* 13: 37–56.

Involuntary Unemployment

John B. Taylor

Keywords

Aggregate demand; Efficiency wages; Frictional unemployment; Incentive wages; Involuntary unemployment; Labour supply; Marginal rate of substitution; Matching; Minimum wages; Natural rate of unemployment; Nominal wages; Non-accelerating inflation rate of unemployment (NAIRU); Optimal contracts; Real business cycles; Real wages; Search theory; Staggered wage setting; Sticky wages; Unemployment; Unemployment measurement; Wage differentials; Wage rigidity

JEL Classifications

J6

The most common and analytically useful definition of involuntary unemployment is based on the labour supply curve: if workers are off the labour supply curve – so that there is an excess supply of labour at the current real wage – then, by definition, there is involuntary unemployment. The amount of involuntary unemployment is equal to the amount of excess labour supply. If workers are on the labour supply curve, then, by definition, there is no involuntary unemployment. One could analogously define involuntary overemployment as a situation of insufficient supply of labour at the prevailing real wage (as may occur during

wartime with wage and price controls), but the term is seldom used.

In a static, deterministic, utility maximization framework, the labour supply curve is simply the set of real wage and employment pairs for which the marginal rate of substitution of income for leisure is equal to the current real wage. Hence, involuntary unemployment can be equivalently defined using the utility function: if the real wage is greater than the marginal rate of substitution of income for leisure, then, by definition, there is involuntary unemployment. If the marginal rate of substitution of income for leisure is equal to the real wage, then there is no involuntary unemployment.

Historical Examples of Usage

This definition of involuntary unemployment is very close to that used by Keynes (1936). In Chapter 2 of the *General Theory*, Keynes writes '. . . the equality of the real wage to the marginal disutility of employment . . . corresponds to the absence of "involuntary" unemployment' (p. 15). (Keynes makes the simplification that the marginal utility of income is constant, so that the marginal disutility of employment is the same as the marginal rate of substitution of income for leisure.) Keynes excluded frictional unemployment from involuntary unemployment. However, it is important to note the Keynes also excluded unemployment 'due to the refusal or inability of a unit of labour, as a result of legislation or social practices or of a combination for collective bargaining or of a slow response to change or of mere human obstinacy, to accept a reward corresponding to the value of the product attributable to its marginal productivity' (Keynes 1936, p. 6). Thus, Keynes chose to exclude union wage differentials as well as minimum wage legislation as sources of involuntary unemployment. Clearly, Keynes wanted to focus on a particular type of involuntary unemployment.

Patinkin (1965, ch. 13) also used the static labour supply definition in his well-known analysis of involuntary unemployment:

> The norm of reference to be used in defining involuntary unemployment is the supply curve for labor . . . as long as workers are 'on their labor supply curve' – that is, as long as they succeed in selling all the labor they want to at the prevailing real wage rate – a state of full employment will be said to exist in the economy. (pp. 314–15)

Although Keynes developed and emphasized the idea of involuntary unemployment much more than economists had done before, the above definition based on the labour supply curve predates Keynes writings. In fact it was used by the 'classical' economists. For example, in 1914 Pigou proposed measuring involuntary unemployment of a group of persons by the number of hours' work by which employment '. . . falls short of the number of hours' work that these persons would have been willing to provide at the current rate of wages under current conditions of employment' (see Casson 1983, p. 39). According to Keynes, however, classical theories (such as Pigou's) did not admit the possibility of involuntary unemployment. Unemployment of a particular group caused by union wage differentials or minimum wage legislation was admitted by the classical theory, but as mentioned above Keynes chose to classify this as voluntary.

Criticisms of the Definition of Involuntary Unemployment

Despite the analytical simplicity of the above definition based on labour supply, the term involuntary unemployment has resulted in many critiques and controversies.

One of the criticisms stems from simple conflicts between the above technical definition and everyday non-technical usage of the term involuntary. For example, Fellner (1976) wrote, '. . . distinguishing elements of voluntariness from elements of involuntariness in the unemployment problem is a hopeless endeavour . . .' (p. 134) and that 'Keynes' definition is unhelpful and so are all variants inspired by that definition' (p. 53). Fellner and others have been concerned that one can never determine the intentions of a given unemployed person so that the broad

classification of unemployment into involuntary and voluntary is meaningless.

Although the many connotations of the term involuntary may cause semantic difficulties (as may other concepts in economics such as 'rational' or 'marginal'), focusing on the technical definition given above would seem to avoid these difficulties.

A second criticism arises in the practical use of the concept of involuntary unemployment for public policy. From the above definition, one criterion of good macroeconomic performance would be zero, or very small, involuntary unemployment. (Strictly speaking, this is true only if the measured real wage is equal to the marginal productivity of labour, an equality that might not hold if optimal contracts of the type described below are important in the economy.) Since government unemployment statistics are commonly taken as an indicator of economic performance, one might hope that measured unemployment could be related to the concept of involuntary unemployment. However, this is very difficult and any attempt is bound to be criticized. Government unemployment statistics typically attempt to measure the number of unemployed who are looking for work, but who have not yet found work. However, aside from the problem of determining whether someone is looking for work, or how intensively, unemployment statistics obviously include frictional unemployment and other types of unemployment that would not be included as involuntary according to the above definition. Even in a condition of relatively full employment, there exists some 'normal' unemployment, which government statistics need to be corrected for. Milton Friedman (1968) used the term 'natural' unemployment for the amount of unemployment that would exist, without excess supply, in equilibrium after wages and prices have adjusted. Another concept of normal unemployment is the non-accelerating inflation rate of unemployment (NAIRU), defined as the amount of unemployment that would exist when there is no tendency for wage or price inflation to rise or fall. Measuring the 'natural' rate or NAIRU in practice entails looking for an unemployment rate for which inflationary pressures are small and adjusting this rate for known changes in the demographic characteristics of the labour force. The natural rate of unemployment is not a constant, however, and these measurements have considerable error. Nevertheless, a practical alternative to involuntary unemployment as a measure of economic performance is the difference between the actual unemployment rate and the natural unemployment rate. For policy purposes, this may serve as a reasonably close approximation to involuntary unemployment, but clearly it is a different concept. In particular, note that this measure can be negative, as when the unemployment rate falls below the natural rate in boom times. Fellner (1976) suggested focusing on this measure and hence on inflation stability, rather than on involuntary unemployment, and he argued that demand management (monetary and fiscal policy) should promote the maximum amount of employment that can be achieved without inflation instability. This measure is also the criterion used in stabilization studies that characterize a macroeconomic trade-off in terms of the fluctuations of unemployment about the natural rate versus the fluctuations in inflation (see Taylor 1980).

A third reason for criticism of the term involuntary unemployment is that the standard definition is essentially static and deterministic. In fact, the static, deterministic labour supply and demand model does not admit an explicit theory of frictional or natural unemployment. Without such a model it is difficult even to discuss whether a given level of unemployment is voluntary or optimal or not.

Research on the microfoundations of unemployment (see for example Phelps et al. 1970), had as a major goal the development of a model of equilibrium unemployment – using search and matching theory. Some search models generated unemployment that was Pareto optimal (see Lucas and Prescott 1974, for example), but others included trading externalities and generated unemployment which could be non-optimal (see Diamond 1982, for example). While not yet definitive, at the least this research shows that for many public policy questions it is necessary to go beyond the simplest model of labour supply, and

thereby beyond the simple definition of involuntary unemployment.

In the *General Theory* Keynes presented a more convoluted definition of involuntary unemployment, and this has been a fourth source of controversy. According to Keynes (1936, p. 15),

> Men are involuntarily unemployed if, in the event of a small rise in the price of wage-goods relatively to the money-wage, both the aggregate supply of labour willing to work for the current money-wage and the aggregate demand for it at that wage would be greater than the existing volume of employment.

One can clearly envisage a point off the labour supply curve from this definition. However, there is much more. Embedded in the definition of involuntary unemployment are some of Keynes's other ideas that were part of his *theory* of involuntary unemployment, but logically distinct from the *definition* of involuntary unemployment. Within the definition it is noted that workers would be willing and able to have a reduction in their real wage (and still increase their work) if it occurred through an increase in the price level, but not if it occurred through a decline in the nominal wage. This 'stickiness' of nominal wages, which is generated as part of the market mechanism, is of course crucial to Keynes's theory. Also embedded in the definition is the assumption that firms are in their labour demand curve, so that a lower real wage would stimulate unemployment, an idea that is much less crucial for Keynes's ideas, as Leijonhufvud (1968) has emphasized. Why did Keynes emphasize this convoluted definition of involuntary unemployment? It seems clear that he wanted to highlight the crucial difference between his theory of unemployment and what he called classical theory. This difference centred on the inability, given the way labour markets and the whole economy interact, of individual workers to reduce unemployment simply by reducing nominal wages. As indicated above, Pigou based the definition of involuntary unemployment on the labour supply curve in much the same way that Keynes did, but the classical reason for its existence – simply that real wages were too high – was much different from the theory of deficient aggregate demand put forth by Keynes. In retrospect Keynes would have added clarity to his discussion by unbundling his theory and his definition of involuntary unemployment.

Implications of Recent Technical Research for the Concept of Involuntary Unemployment

Five research developments since the 1960s have had great relevance for the concept of involuntary unemployment: equilibrium macroeconomics, optimal contract theory, disequilibrium macroeconomics, efficiency or incentive wage theory, and staggered- wage setting theory. However, this relevance must be inferred from the research, because the term involuntary unemployment is seldom used explicitly, and perhaps avoided by many recent researchers.

Equilibrium Macroeconomics

One strand of research macroeconomics has established a strategy of trying to explain the observed fluctuations in unemployment by equilibrium models in which workers are always on their labour supply curves. Wages and prices are perfectly flexible, and all markets clear in these models. Lucas and Rapping (1969) and Kydland and Prescott (1982) represent some of the seminal work in this strand of research. Clearly if these models turn out to be successful and to dominate other models, then the idea of involuntary unemployment would become useless for macroeconomics. Shifts of the labour supply curve – caused by intertemporal substitution of labour supply in response to temporary actual or perceived fluctuations in the real wage – are the main source of employment variability in these models. Research in this area is continuing and branching out into 'real business cycle' theory which ignores monetary factors in the cycle altogether. It appears, however, that a very high labour supply elasticity – by the standards of recent microeconomic empirical research (see MaCurdy 1981) – is required for these models to be able to explain the observed fluctuations in employment.

Optimal Contract Theory

Studies by Azariadis (1975), Baily (1974) and others attempted to explain why involuntary unemployment would arise when there exist optimal contracts between firms and workers stipulating for fixed wage payments. However, when firms and workers have equal access to information, these studies have shown that, in the relevant sense, involuntary unemployment does not exist despite the fixed wage bill. In these optimal contract models the marginal rate of substitution of income for leisure is equal to the marginal productivity of labour – the condition for the optimality – in all possible states. Although workers are off their labour supply curve *ex post* (since the real wage is not necessarily equal to the marginal rate of substitution), this discrepancy has no welfare significance. Models in which firms have more information than workers about the nature of the shock can lead to a breakdown in the marginal conditions for optimality, but unless firms are more risk averse than workers the result is involuntary *over*-employment: the marginal productivity of labour is less than the marginal rate of substitution of income for leisure (see Green and Kahn 1983; Grossman and Hart 1983). Viewed as an attempt to explain involuntary unemployment this research, therefore, has been unsuccessful. Taken literally, it shows that much of the unemployment that may have appeared as involuntary is, in fact, voluntary or at least efficient!

Disequilibrium Theory

Malinvaud's (1977) careful examination of fixprice multimarket equilibria, following the tradition of Clower (1965) and Barro and Grossman (1971), has greatly helped to clarify the conceptual difference between Keynes's explanation of involuntary unemployment due to insufficient aggregate demand (where firms are constrained in product markets), and the classical unemployment associated with the real wage being too high (where firms are not constrained in product markets). This research also has had considerable policy relevance in the early 1980s because the high rates of unemployment in western Europe were diagnosed as classical rather than Keynesian by many economists.

Efficiency or Incentive Wages

Calvo (1979) and others have argued that involuntary unemployment can occur because high wages must be paid to give workers the incentive to work hard, to be productive, and not to shirk. As firms attempt to bid up their wages relative to other firms, an equilibrium is reached with all firms paying more than the wage in the absence of incentive effects and with involuntary unemployment: an excess supply of labour with unemployed workers willing to work at the going wage. This type of unemployment is not of the deficient demand type emphasized by Keynes, and given Keynes's willingness to lump other minimum wage unemployment in with frictional unemployment, it is likely that Keynes would have classified this type of unemployment as voluntary. Incentive wages would increase the normal unemployment (natural or NAIRU) rate, but there is little empirical evidence of how quantitatively important the effect is.

Staggered-Wage Setting Theory

In these models (see Taylor 1980, for example), wages are set with an aim to maintain relative wages unless there is a reason for relative wages to adjust. This relative wage setting leads average nominal wages to adjust with a lag described by a predictable dynamics to changes in demand. In these models prices are set as a markup over wages, and for this reason aggregate prices are almost as sticky as nominal wages. Combined with an elementary model of aggregate demand and an aggregate demand policy that does not fully accommodate inflation, these models are designed to be compared directly with the data and in fact lead to fluctuations in unemployment which have features similar to the real world. The unemployment in these models comes close to the usual definition of involuntary unemployment, but since explaining empirical regularities is a primary objective, unemployment enters the model directly as the deviation of unemployment from the natural rate – a more readily measurable quantity than involuntary unemployment. These models show that wage rigidities need not be very long to generate the type of fluctuations in unemployment that characterize the business cycle.

Like the equilibrium models discussed above, and unlike the other three research developments described above, these models are dynamic and can therefore be directly tested against time series data.

Although there has been a tendency for much recent research to avoid the term involuntary unemployment, and instead to define unemployment as appropriate to the theoretical or empirical objectives of the research itself, the term involuntary unemployment will probably continue to be used. Despite the criticism and controversy discussed above there is little harm in this usage, as long as the technical definition is emphasized. Its usage may encourage researchers to point out the connection of new results to past achievements.

See Also

- ▶ Implicit Contracts
- ▶ Natural Rate of Unemployment
- ▶ Search Theory
- ▶ Unemployment

Bibliography

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83: 1183–1202.

Baily, M.N. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41: 37–50.

Barro, R.J., and H. Grossman. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61: 82–93.

Calvo, G. 1979. Quasi-Walrasian theories of unemployment. *American Economic Review: Papers and Proceedings* 69: 102–107.

Casson, M. 1983. *Economics of unemployment*. Oxford: Martin Robertson.

Clower, R. 1965. The Keynesian counter-revolution: A theoretical appraisal. In *The theory of interest rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Diamond, P. 1982. Aggregate demand management in search equilibrium. *Journal of Political Economy* 90: 881–894.

Fellner, W. 1976. *Towards a reconstruction of macroeconomics: Problems of theory and policy*. Washington, DC: American Enterprise Institute.

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

Green, J., and C. Kahn. 1983. Wage employment contracts. *Quarterly Journal of Economics* 98(Supplement): 173–187.

Grossman, S.J., and O. Hart. 1983. Implicit contracts, moral hazard, and unemployment. *American Economic Review* 71: 301–307.

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

Kydland, F., and E.C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. New York: Oxford University Press.

Lucas, R.E. Jr., and E.C. Prescott. 1974. Equilibrium search and unemployment. *Journal of Economic Theory* 7: 188–209.

Lucas, R.E. Jr., and L. Rapping. 1969. Real wages, employment, and inflation. *Journal of Political Economy* 77: 721–754.

MaCurdy, T.E. 1981. An empirical model of labour supply in a life cycle setting. *Journal of Political Economy* 89: 1059–1085.

Malinvaud, E. 1977. *The theory of unemployment reconsidered*. Oxford: Basil Blackwell.

Patinkin, D. 1965. *Money interest and prices*, 2nd ed. New York: Harper & Row.

Phelps, E.S., et al. (eds.). 1970. Microeconomic foundations of employment and inflation theory. New York: W.W. Norton.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.

IQ and National Productivity

Garett Jones

Abstract

A recent line of research in economics and psychology hypothesizes that differences in national average intelligence, proxied by IQ tests, are important drivers of national economic outcomes. Cross-country regressions, while showing a robust IQ-growth relationship, cannot fully test this hypothesis. Thus, recent work explores the micro-foundations of the IQ-productivity relationship. The well-identified psychological relationship between IQ and patience implies higher savings rates and higher folk theorem-driven institutional quality in high average IQ countries.

Experiments indicate that intelligence predicts greater pro-social behavior in public goods and prisoner's dilemma games, supporting the hypothesis that high national average IQ causes higher institutional quality. High average IQ countries also have higher savings intensity by a variety of measures. Other possible IQ-productivity channels are discussed, as are possible environmental causes of differences in national average IQ.

Keywords

Cognitive ability; Economic growth; Education; Human capital; Time preference; Institutions; Cooperation; prisoner's dilemma; IQ; Intelligence; Capital; Strategic complementarities; Intelligence quotient; IQ tests; GDP; Average worker productivity; Cognitive skill

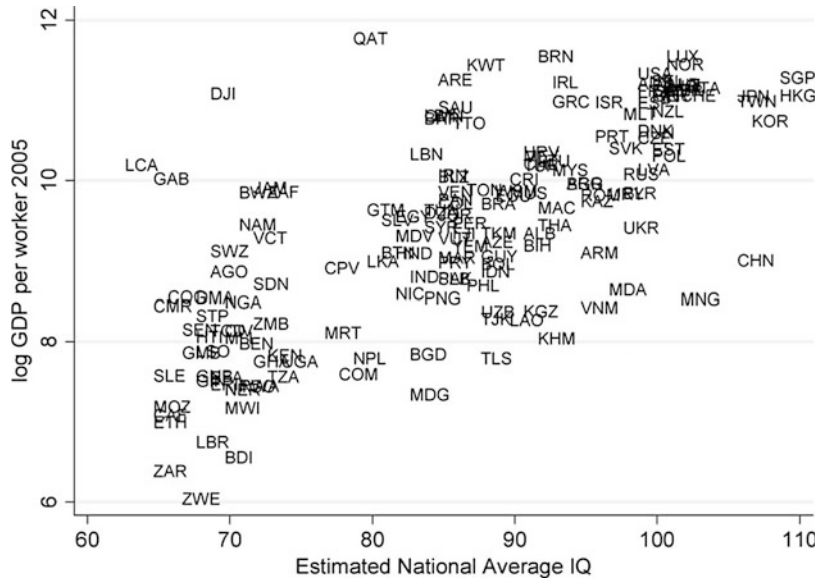
JEL Classifications

D91; I15; J24; O43

In recent years, some economists and psychologists have proposed that the average level of intelligence in a country – measured by conventional IQ tests – is an important independent driver of economic outcomes. As psychologists have known for decades, average IQ scores differ when given to large samples in different nations, and recent estimates indicate national average IQ correlates 0.7 with log GDP per capita (Fig. 1).

The macroeconomic question of interest is why IQ, which has a modest relationship with individual wages within a country, has such a strong relationship with average worker productivity across countries. When IQ is normalised in conventional IQ points (UK mean = 100, standard deviation = 15), one IQ point is associated with approximately 1% higher wages, but 6 to 7% higher national productivity. While higher national productivity almost certainly raises IQ for the poorest countries, this reverse causation may not be the whole story, and indeed channels running from income to IQ have been widely studied already (Jones and Schneider 2006, 2010, and citations therein.). This article focuses instead on potential cognitive spillovers, channels through which national IQ could have payoffs to nations that only indirectly reward high-IQ individuals. These possible channels include better-informed voters, more cooperative workers and political elites, more saving (in a world of capital frictions) and access to higher-quality production functions (in a world of complementarities to worker quality).

A brief summary of IQ testing and of the possible drivers of national IQ differences is provided



IQ and National Productivity, Fig. 1 National average IQ and year 2000 GDP per worker (Notes: Y-axis shows GDP per worker in logarithmic scale. The sample covers 164 countries. Pearson and Spearman correlation both equal 0.7. One IQ point predicts 7.7% higher GDP per worker. Source: Lynn and Vanhanen (2006) and Penn World Tables 7.0 for IQ and GDP data, respectively)

below; the discussion includes possible methods of raising national average IQ.

The essay proceeds as follows. We begin by summarising modern psychometric estimates of intelligence and the major databases of national average IQ. We then discuss the labour literature on IQ and wages, followed by growth regressions that control for national IQ. Two channels – patience and skill complementarities – by which intelligence might influence national GDP per capita are then considered, before turning to institutional channels. Methods that could raise national average IQ, including environmental, nutritional, and health interventions, educational interventions, and immigration policies are discussed, and finally we look at areas for future work and draw conclusions.

Measuring Intelligence

What is intelligence? Can it be measured? The answers to the first question range from 'intelligence is what intelligence tests test' (Peak and Boring 1926, p. 71) to the more useful statement that intelligence is a model of mental ability built around 'the empirical fact that all mental abilities are positively correlated' (Jensen 1998, p. 45). In other words, people who are better at mathematical reasoning tend to do better than average on trivia tests, whereas people who are worse than average at pattern-finding tend to be worse than average at vocabulary tests or memorizing long lists of numbers. The precise value of the correlation across mental tasks varies across tasks but in large, diverse samples it is non-negative. (For a candid consensus document on the nature of intelligence authorized by the American Psychological Association, see Neisser et al. (1996).)

Psychologists have found this positive correlation so often that in academic research, intelligence is often operationalized as the '*g* factor', the first principal component from a large battery of mental tests (Jensen 1998, Ch. 3). This one summary statistic, *g*, is often translated into units known as an IQ score. IQ is normed at a mean value of 100 within the UK, and the standard deviation of IQ within the UK is defined as equal to 15 IQ points. In practice, and within this essay, other types of IQ test that use other metrics are converted into the IQ scale with UK mean 100 and standard deviation of 15.

A vast literature across the social sciences has documented the many conditional and unconditional correlates of IQ: within affluent countries, individual IQ scores correlate positively with lifespan, wages and (*sic*) myopia; it correlates negatively with criminality, tendency to smoke and number of motor vehicle accidents (Jensen 1999, Ch. 9). These correlates are well-known. Some less-well-known correlates of individual IQ include *in vivo* brain size measured with MRIs and CT scans (typically +0.3 to +0.4, cf. Wickett et al. 2000), nerve conduction velocity between the eye and the vision centers of the brain (+0.4), and reaction time and inspection times (respectively, speed with which one presses a lighted button and the minimum amount of time one needs to decipher whether a quickly-flashed symbol was, say, an 'I' or an 'L') (Jensen 1998, Ch. 6; Deary 2001). These newer correlates of IQ are much less subject to the criticism that people with high IQs are just people who test well. Instead, they are something more: they are quicker.

Can IQ be measured across countries, even in developing countries? And if so, do these tests have similar real-world reliability to IQ tests given within OECD countries?

The answer to both questions is yes, with some modest grounds for caution. IQ tests have been translated into dozens of languages, and private companies who sell IQ tests to schools, hospitals and firms have often created nation-level standardization samples of 1,000 or more. Although non-psychologists often think of IQ tests as 'pencil and paper tests', in fact the widely used Wechsler IQ tests involve mostly talking with a psychologist and answering her verbal questions; a few subtests involving solving wood block puzzles, assessing pictures or (occasionally) writing some answers with pencil and paper.

Further, IQ tests exist that are entirely non-verbal: the Catell Culture-Fair test, Raven's Progressive Matrices and the Draw-a-Man test are three prominent examples. So any researcher who

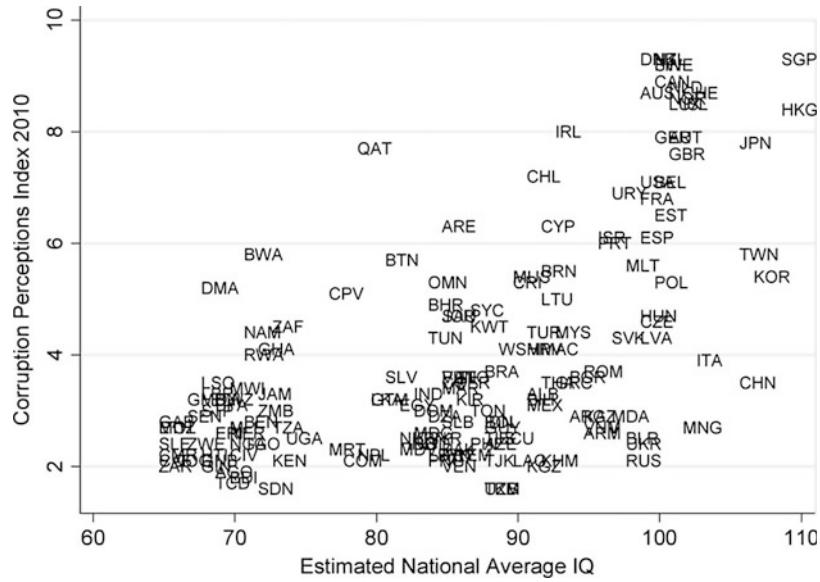
chose to estimate a nation's average IQ score would have a wide variety of tests from which to draw.

The psychologist Richard Lynn and the political scientist Tatu Vanhanen (henceforth LV) assembled two collections of IQ scores by scouring the academic and practitioner literatures for reported IQ in a total of 113 countries (2002, 2006). They included some IQ standardisation samples and some national tests of mathematical ability, but most of the studies they used were 'opportunity samples', studies of an ostensibly typical classroom or school in a particular country. As Jones and Schneider (2010) show, the high-quality samples and opportunity samples are highly correlated, and have a mean absolute deviation of 3.2 IQ points.

Recall that the LV IQ estimates correlate 0.7 with log GDP per capita. Because the LV sample includes many types of IQ test and because LV describe the IQ tests that make up each nation's IQ estimate, Jones and Schneider (2010) were able to show that this correlation holds if one uses only IQ estimates from non-verbal IQ tests. The correlation between Ravens IQ and log GDP per capita is between 0.9 and 0.7, depending on the form of the test; the Ravens was the only single test used often enough to calculate test-specific correlations. And regardless of the type of IQ test used, rank order across countries is little-affected.

LV used these data to create estimates of national average IQ; theirs were the first databases of national average IQ, but not the last. When LV had multiple plausibly representative IQ estimates for a country, they took the mean (2002) or the (likely superior) median (2006) from across the studies. In their 2006 dataset, they have data from 113 countries, and for most countries they have more than one study to draw upon. The global mean IQ (unweighted by population) is 90, 2/3 of a UK standard deviation below the mean, and the standard deviation across countries is 11 IQ points. In recent work, Rindermann (2007a, b), Rindermann and Thompson (2011) and Lynn and Meisenberg (2010a) have created new average national IQ estimates using more rigorous methods but for fewer countries; since their estimates correlate strongly with the larger LV 2006 sample, we largely use the latter in this essay (LV also interpolate national IQ estimates for countries that lacked IQ scores: In LV (2006) they present evidence that their LV (2002) interpolations were reliable. Figures 1 and 2 use both interpolated and actual IQ estimates; correlations are unaffected when interpolated values are omitted).

One question is whether these IQ measures across countries are reliable: whether they



IQ and National Productivity, Fig. 2 National IQ and corruption (Notes: Higher values on Y-axis indicate lower perceived corruption. The sample covers 165 countries. Pearson and Spearman correlations both equal 0.6. *Source*: Lynn and Vanhanen (2006) and the 2010 Corruption Perceptions Index)

measure differences in the same battery of mental skills across countries as they do within countries. On a variety of measures, one can say that the answer is yes. Leaving aside the purely psychometric measures of cross-cultural IQ validity (a longstanding research area in psychology; see citations in Jensen (1998, Ch. 11)), economists have found that within low average IQ countries, IQ scores have approximately the same relationship with wages as they do in rich countries. In both rich and poor countries, 1 IQ point is associated with between 0.5% and 1.25% higher wages. One study in rural Pakistan using the Ravens IQ test (Alderman et al. 1996) found that 1 IQ point was associated with 0.9% higher wages, very close to the Zax and Rees (2002) estimate of males in Wisconsin; and other examples can be multiplied (Behrman et al. 2004).

Some social scientists have criticized the LV datasets (Volken 2003; Wicherts et al. 2010a and citations therein); early criticisms included claims of one-to two-point errors in recording or interpreting the underlying data. In only one case was a misinterpretation substantial (Equatorial Guinea, mistakenly given an IQ estimate of 59; Lynn has dropped this observation from his most recent update). Random errors in the one-to two-point range are regrettable but almost surely irrelevant for empirical work; and to the extent that they introduce errors in variables, they will understate the true relationship in bivariate regressions (Durbin 1954).

More recent criticisms arose in a series of papers by Wicherts et al. (2009, 2010a, b) focusing solely on LV's sub-Saharan African IQ estimates. These critics note that LV exclude many studies of African IQ from their sample, and include some studies where the researchers reported health problems or enormous irregularities in test administration (for instance, some children taking a test in rural Africa were inexperienced in the use of pencils). Lynn and Meisenberg (2010b) responded to this critique in part by noting that tests preferred by Wicherts sometimes included college student samples or otherwise elite populations, samples likely to be unrepresentative in undeveloped countries; indeed, LV always omitted college-only samples when estimating IQ for rich countries, so their treatment was symmetric across rich and poor. This exchange is highly recommended for insight into how databases are constructed; and Young (2010) is recommended as a parallel reminder of the weaknesses of African GDP data.

On the question of student health, if poor health hurts both measured IQ and the underlying skill the IQ test is designed to measure, then researchers should *hope* that such samples of students are included in a national IQ estimate: while there is certainly real interest in knowing what a nation's average IQ would be if all students had first-world health and nutrition, it is also of great interest to know how student's brains are performing in the world as it currently exists.

In a surprise ending to the dispute between Lynn and Wicherts et al., the latter chose (2010b) to look at individual studies that used only large, nationally representative samples, samples that met all of their quality requirements. In their K-12 samples, the median IQ across a variety of sub-Saharan African countries was 76.5 – about half an intra-UK standard deviation away from Lynn's own estimate of average sub-Saharan African IQ: 70. By either measure, sub-Saharan African nations currently have the lowest average IQs of any region of the world. As Wicherts et al. themselves conclude, '[t]here can be little doubt that Africans average lower IQs than do Westerners' (Wicherts et al. 2010a, p. 17). Wicherts et al. (2010a, p. 17) propose some methods of increasing average African IQ: 'These include improvements in nutrition and health [care], increases [in] educational attainment, improvements in educational practices, urbanization, large-scale dissemination of visual–spatial toys, etc. Although it cannot be precluded that genetic effects play a role in the low IQ performance of Africans, we view environmental circumstances as potentially more relevant to the present-day difference in mean'.

Thus, the academic critics and Lynn agree on the point that is of most interest to economists: IQs differ across countries, and the rank order of the difference has broad agreement. As we shall see below, in applied cross-country work researchers have both Winsorized the data to 80 or 90 or

included sub-Saharan African dummies, partly to take account of the possibility that these scores are inaccurately low. Researchers have taken the Wicherts et al. critiques into account.

IQ and Wages

A routine finding in labor economics is that childhood and adolescent IQ scores in developed countries are positively correlated with adult wages (Bowles et al. 2001; Neal and Johnson 1996, and citations therein). This holds whether or not one controls for education. In a conventional diminishing returns, price-taking setting, the relationship between IQ and wages is the relationship between IQ and the marginal product of labor. Jones and Schneider (2010) use this fact to estimate the microeconomic effect of differences in national IQ on national productivity.

Using their preferred estimate for ease of exposition, 1 IQ point is assumed to cause 1% higher micro-level productivity.

Based on Ramsey and Solow model intuitions, economists might expect that at the national level, in steady state, this would cause an even greater effect on macro-level productivity: IQ raises the marginal product of labor, and since in a Cobb–Douglas production function this is observationally equivalent to an increase in TFP, this would raise the marginal product of capital as well. In steady state, one would expect this to increase the level of capital in order to reduce the marginal product of capital back to its steady state level. All of these channels are correct, but nevertheless conventional theory predicts that for conventional production functions, the micro level IQ–productivity relationship is identical to the macro steady state IQ–productivity relationship.

The micro and steady-state macro effects are identical because the micro relationship among human capital, wages, and productivity already assumes that capital is drawn to highly productive labor. Consider the case of two types of workers with IQ levels IQ_{L} and IQ_{H}, $IQ_{\mathrm{H}} > IQ_{\mathrm{L}}$. The size of each group of workers is normalized to unity. The workers share a fixed homogenous capital stock $K_{\mathrm{L}} + K_{\mathrm{H}} = K$ and use the production function

$$Y = K_L^{1/3} IQ_L^{2/3} + K_H^{1/3} IQ_H^{2/3}.$$

If capital can flow freely and the representative firm is a price-taker, then capital will flow toward the high IQ workers until the marginal product of capital is equalized across the two categories of workers. In equilibrium, workers with 10% higher IQ will have 10% more capital, because they provide 10% more effective labor. One can place quotation marks in the previous sentence around either ‘more effective’ or ‘effective labor’: the management science interpretation is different, but the implications for the capital–labor ratio are the same. Thus, micro-level cross-section wage regressions include the endogenous effect of physical capital’s attraction to workers with higher levels of human capital.

One can use the cross-sectional relationship between IQ and wages within a country to test the validity of the LV IQ measures. In an ideal test, one would want to randomly select workers from different countries and place them into a new country, wait a few years, and then measure their wages. One could then determine whether the average immigrant from a country with an LV IQ 10 points lower than her new country earns 10% less than the average person in the new country. This experiment would hold institutions, capital and many other features constant, and only vary the country of origin of the experimental subjects.

Such an experiment is impossible and undesirable in its pure form, but immigrants to the USA provide a useful approximation. Using Hendricks’s (2002) data on wages of immigrants to the USA, Jones and Schneider (2010) regress the average income of immigrants from each country (whether or not adjusted for education) on that nation’s LV IQ. They find approximately the same 1:1 IQ/wage relationship that others find in the labor literature: immigrants from higher LV IQ nations earn modestly more after arrival in the USA than immigrants from lower LV IQ nations. A one-point increase in national average IQ predicts an approximate 1% increase in average income of immigrants from that country. In a simple calibration Jones and Schneider find that this private marginal product of labor channel can

explain approximately 1/6th of the cross-country variation in log productivity per worker. Workers from higher LV IQ countries are typically more productive, although this private productivity channel is likely far from the whole story.

National IQ in Growth Regressions

Growth regressions also support the hypothesis that higher LV IQ causes better economic performance. LV themselves ran bivariate correlations and one-and two-variable growth and level regressions, always finding a strong relationship between national average IQ and the outcome of interest. Weede and Kampf (2002) ran more conventional Mankiw–Romer–Weil/Barro-style cross-section growth regressions, controlling for institutional quality, schooling and starting GDP per capita; again, national IQ was a reliable growth predictor. While some regressions had been reported, the question of LV IQ's overall robustness was unclear.

Jones and Schneider (2006, henceforth J/S) answered the question of LV IQ's robustness by including (*inter alia*) 455 cross-section growth regressions using the dataset and combinations of the control variables from Sala-i-Martin et al. (2004, henceforth SDM). J/S only included the 18 growth regressors that were robust in SDM's Bayesian Averaging of Classical Estimates exercise. They included geographic dummies, years open to trade and ethnolinguistic fractionalization among other controls. Their regressions included three fixed controls (the three most robust SDM regressors: log GDP per capita in 1960, primary schooling in 1960 and the price of investment goods) and all 455 possible permutations of the remaining 15 robust controls taken three at a time. Thus seven controls were included in every regression. National average IQ was statistically significant at the 1% level in *every* regression.

J/S provide additional Bayesian model averaging robustness tests that accord with this result. Ram (2007) also found that national IQ was a statistically significant growth regressor in multiple specifications. Using the structural equation methods common in psychology, Rindermann and Thompson (2011) have found a reliable positive relationship among national average cognitive skills, good pro-market institutions, and good economic performance. Notably, that paper includes a separate estimate of the cognitive skills of the highest-scoring 5% of the population in his OECD-heavy sample; they find that the skills of the top 5% have disproportionate predictive power for good outcomes.

J/S use their mean IQ growth regression coefficient to calculate the predicted steady-state relationship between LV IQ and GDP per capita. This calculation, derived from Jones (2000) and Barro and Sala-i-Martin (2001, pp. 466ff.), is both practical and rarely used. Recall the conventional cross-sectional growth regression, where y_i is GDP per capita for country i at the beginning of the time period, $\Delta\ln(y_i)$ is the log change in GDP per capita over the sample period, β is the speed of convergence to steady state (typically found to be 2% per year (Barro and Sala-i-Martin 2001, pp. 496, 521; for an early critique see Quah 1996)), X_i is a column vector of other controls, and θ' is the row vector of coefficients corresponding to those controls:

$$\Delta\ln(y_i) = \gamma - \beta\ln(y_i) + \theta' X_i + \varepsilon_i$$

If we assume that technology grows at an exogenous rate (a conventional motivation for treating γ as exogenous to a particular country is that the growth rate of useful technical knowledge is overwhelmingly external to any one country) and appropriately demean the other variables, then γ is the steady-state growth rate of GDP per capita, conventionally considered 2% per year over the past century. Under the Solow-style assumption of conditional convergence, all steady-state growth in per capita GDP is caused by γ: so any growth seen over the sample period greater or less than γ is caused by convergence to a nation's steady state path of GDP per capita. This suggests the following transformation:

$$\Delta\ln(y_i) = \gamma + \beta[\lambda' X_i - \ln(y_i)] + \varepsilon_i$$

where $\lambda = \theta/\beta$. By factoring out a β from the coefficient on the so-called growth regressors,

we see that under the assumption of conditional convergence 'growth regressors' are actually steady-state log-level regressors. Therefore the coefficients θ on these growth regressors are actually coefficients for the steady-state log-level effect λ multiplied by the rate of convergence, β. The term in square brackets has a straightforward interpretation: it is the gap between starting log GDP per capita and steady-state log GDP per capita. (Strictly speaking, $\theta' X_i$ is log steady-state GDP per capita as of the end of the sample period: given the exogenous growth rate, each nation's log steady-state GDP per capita increases by γ every period.) The gap between the two closes at rate β per year.

Jones and Schneider find that 1 IQ point is associated with slightly more than 0.1% faster annual GDP growth; given their β estimate of slightly less than 2, they estimate that 1 IQ point predicts 6% higher steady-state GDP per capita. As noted already, this is at least six times greater than most estimates of the micro-level relationship between IQ and individual productivity, and it is consistent with the hypothesis that IQ has positive spillovers. The next two sections discuss what some of those cognitive spillovers might be.

Skill Complementarities and Patience

This section considers two channels for IQ spillovers: complementarities to worker skill and the well-identified link between patience and intelligence.

Kremer (1993) notes that much of modern production is fragile: He discusses the explosion of the space shuttle *Challenger*, destroyed by the failure of a single O-ring, a band that sealed the burning rocket engine. Less tragically, and more routinely, many production processes have such 'weakest link' elements, where many workers toil on a project, and where failure at any one step of the production process can destroy the value of the whole. As Kremer notes, clothing with minor flaws is sold at steep discounts and computer chips with the smallest flaws are unusable.

In weakest link settings, Kremer shows that it is privately rational and socially efficient for workers to sort across firms by quality. To take the simplest example, assume two types of worker of quality q_{H} and q_{L}, with effective labour of $q_{\mathrm{L}} = 0.5q_{\mathrm{H}}$. The production process has two steps and output is multiplicative in worker quality: $Y = q \times q$. This production function could represent, as Kremer notes, the process of making a vase, where any worker has some probability $1-q$ of dropping the vase (the probability of completing a particular vase is then q^2); or where an error at any step in the production process would reduce the value by some fraction q.

Given a fixed supply of workers and equal amounts of both types of workers, it is efficient to sort workers into firms where all workers are of the same quality:

$$q_H^2 + q_L^2 = 1.25q_H^2 > 2q_H q_L = q_H^2.$$

Within a given firm that had both types of workers, the firm would voluntarily sort the workers to maximise output. Kremer shows that this result is generalisable to decentralised economies with physical capital stocks and large varieties of worker skill: If production functions have complementarities to skill, market forces will tend to sort workers into firms by skill level.

Kremer draws on this finding to help explain why rich countries tend to have higher levels of human capital: higher average levels of worker skill open the door to using more advanced technologies (which may demand longer production chains), while nations with lower skill levels will be able to produce little output, since long production chains dramatically increase the probability of a value-destroying error.

The Kremer model thus provides a new reason for believing that returns to human capital may be large; but how could one reconcile the hypothesis that human capital returns are large across countries with the routine finding that returns to human capital are modest within countries? The model of Jones (2010) provides one resolution: the model proposes that there are two production technologies available in each country, a Kremer-style O-ring technology and a diminishing returns to

scale, Cobb–Douglas 'Foolproof' technology that works according to the conventional model of Part II. (The diminishing returns in the Foolproof sector reflect multiple ways in which non-O-ring technologies are less productive as they expand as a fraction of the economy: it can represent nontradable personal services, demand for which is limited; it could represent the use of well-understood production processes that were perhaps once O-ring in nature but are now relatively 'Foolproof' (such as the production of aspirin); and it could represent traditional, nonscalable agricultural and manufacturing methods.) In this Cobb–Douglas sector, workers of different skill levels can readily work together in the same production process, and average skill level is a sufficient index of worker quality. In the Foolproof sector, workers that are 1% lower in average quality might early only 1% less; whereas in the O-ring sector, 1% average lower worker quality causes a much larger decline in output and hence wages.

Consider the case of two types of workers, again q_H and q_L: high-skilled workers voluntarily sort between the two sectors until wages for q_H workers equalize. The model's key result is that as long as there are not too many lower-skilled workers, the q_L workers will voluntarily sort into the Cobb–Douglas sector, producing slightly less output and earning only slightly lower wages than other, higher skilled workers in the same country. The q_L workers would not want to use the O-ring technology since they are much less productive with that technology; lower quality workers are poor substitutes in the O-ring sector but good substitutes in the Foolproof sector.

But a nation of q_L workers would produce little with the O-ring production technology; they would likely crowd into the Cobb–Douglas sector, producing little indeed, though still more than if they used the O-ring technology. The Foolproof sector is appealing to low-skilled workers when there are few people in it; because it faces diminishing returns (or limited demand for Foolproof goods and services), a nation of q_L workers in the Foolproof sector will have very low average productivity. Returns to human capital will be small within countries but large across countries. The model is thus consistent with the empirical observation that IQ has a strong relationship with cross-country productivity but a weaker relationship with intra-country productivity.

While the O-ring/Foolproof model matches this fact, further work can investigate whether other implications of the theory hold true: do low-skilled and low-IQ workers take on more complex, delicate tasks when living in nations with low average IQ? Do high-skilled, high-IQ workers take on more mundane tasks when working in high-average IQ countries? Within a country, are new technologies massively more productive when used solely by higher-IQ workers? Production functions, so widely used in economics, are still under-tested in empirical work.

Another link between IQ and national productivity is driven by IQ's reliable correlation with patience. In intertemporal optimising models of national economies, the rate of time preference is always a key parameter, one that influences long-run interest rates, investment, and the capital stock. Growth economists typically assume the rate of time preference is identical across countries. Is this assumption tenable? Psychological research and behavioural economics research combined with LV IQ estimates suggest the answer is no. And recently, Banerjee and Duflo (2011) have written that reduced willingness to delay gratification may be of first-order importance in explaining global poverty: 'the poor... often behave as if they think that any change that is significant enough to be worth sacrificing for will simply take too long. This could explain why they focus on the here and now...'.

Psychologists have known for decades that patience and IQ are almost always positively correlated. Shamosh and Gray (2008) survey this literature, finding that in 23 out of 26 experimental studies, high IQ individuals are more likely to delay gratification.

Shamosh and Gray suggest one channel through which intelligence could directly cause patience: through the ability to keep multiple facts simultaneously in one's mind. One strong correlate of overall intelligence – indeed, one subtest of some IQ tests – is memory span: the quantity of numbers or letters that can a person

can recall a few moments after hearing them. Since considering the opportunity cost of consuming now versus later requires keeping four hypothetical situations in mind (consuming vs. not consuming now; not consuming vs. consuming later), memory span provides one cognitive foundation for the IQ–patience relationship. Further work can investigate other possible channels.

In one well-known study of delayed gratification Mischel et al. (1972), the experimenter gave a 4–6 year-old child a marshmallow, and then told the child that he was going to leave the room. He told the child that if she waited until he returned to eat the marshmallow, the child would get a second marshmallow. The experimenter then waited long enough that almost all children eventually ate the marshmallow (or other treat); minutes until marshmallow eaten was then recorded as the key experimental outcome.

Children used many innovative methods to avoid thinking about the marshmallow on the table in front of them, such as 'covering their eyes with their hands or talking to themselves' so they would think less about the marshmallow (p. 205). These innovations are suggestive of a link between delayed gratification and intelligence. And the evidence supports such a suggestion: in a 1990 follow-up of the adolescent behavior of these same test subjects, Shoda et al. found that children who waited longer before eating the marshmallow had higher SAT verbal ($\rho = 0.42$, $p < 0.05$), and SAT quantitative ($\rho = 0.57$, $p < 0.001$) scores.

Since children's differences in waiting time are measured in mere *minutes*, then any attempt to convert this study's results into a parameter linking SAT (and its strong correlate, IQ – cf. *inter alia* Frey and Detterman (2004) and Beaujean et al. (2006)) to the *annual* rate of time preference would involve astronomical numbers. With half a million minutes per year, any IQ-delay finding from such an experiment extrapolated to the annual β or ρ parameters familiar from growth economics would predict that low average IQ countries would have negligible savings. In the original study, Mischel et al. found that older children waited longer; this suggests that this is not an age-invariant parameter. While the IQ–patience relationship is well-documented, economists will have to search further for a parameter relevant for national economies.

Two recent studies by economists have provided evidence that among adults, the IQ–patience relationship *can* be mapped into the familiar space of choices over long time periods. Dohmen et al. (2010) using a sample of German adults find that in both hypothetical and actual choices of money now versus a year from now, a one intra-US standard deviation increase in cognitive skill is associated with a decline in the discount factor. Further evidence comes from the US peace dividend of the early 1990s: when the US military downsized, it offered enlisted personnel who wanted to separate early the option of an immediate lump-sum payment or an attractive annuity with an internal rate of return greater than 17%. Even controlling for income, years served, age, education and many other factors, scores on an enlisted person's Armed Forces Qualifying Test was a statistically significant, correctly signed predictor of one's likelihood of accepting the attractive annuity (Warner and Pleeter 2001).

Jones and Podemska (2010) convert these estimates of the relationship between cognitive skill and time preference into a parameter, $d\rho/d(IQ)$. Their benchmark estimate is that one IQ point lowers the discount rate by five basis points. They then use that data in a conventional Ramsey growth model. In a closed economy, differences in national IQ would predict less investment and hence lower steady-state capital–output ratios; Jones and Podemska show that indeed high IQ countries tend to have higher capital–output ratios and higher rates of savings. (The empirical relationship between national IQ and savings, or national IQ and capital intensity, is quantitatively larger than one would expect from a simple Ramsey growth model when estimated in logs; it is close to the predicted relationship when estimated in levels. Peer effects on saving deserve attention as one possible reason for a stronger country-level relationship; Maurer and Meier (2008) find moderate peer effects on individual consumption spending using the Panel Study of Income Dynamics.)

This is an IQ externality because the capital stock with which one works and from which one earns interest is determined by the average IQ of one's national compatriots; if one were permitted to move to a higher-IQ country, one would be able to work with a larger capital stock through no effort of one's own. The missing market here is the market for global labour (or equivalently, frictionless global capital flows).

In an economy fully open to capital flows, Barro and Sala-i-Martin (2001, pp. 164–5) describe the quite extreme steady state. In a purely theoretical discussion, they rank countries by order of time preference, denoting Country 1 as the most patient. Their theoretical result is stark:

> Asymptotically, Country 1 owns all the wealth... [all] claims on capital and the present value of the wage income in all countries... All other countries own a negligible amount (per unit of effective labor) in the long run.

The reason for their result? Because the market interest rate – identical around the world – will eventually be set by the most patient country, and less-patient nations will voluntarily borrow money at that interest rate to consume more than their income until they have promised their entire future income stream (minus an epsilon amount, under the Inada conditions) to repay the debt.

Because this conclusion is considered unrealistic, growth economists typically assume that there are frictions that keep any one nation from promising its entire future income stream as collateral. But as barriers to international finance have fallen in recent decades, then one might expect the world to move, if not entirely to the Barro/Sala-i-Martin steady state, at least in that direction.

Jones and Podemska (2010) claim that for many countries, holdings of US Treasuries, a liquid form of wealth, are one indicator of whether a nation is building up its stock of global savings. Omitting a few offshore banking havens and OPEC countries, they find that high LV IQ countries hold a disproportionate share of US Treasuries as a share of their nation's GDP: high LV IQ countries have high Treasury/GDP ratios, a result that holds after controlling for GDP per capita.

The Treasury/GDP ratio is an imperfect index of global saving; they are gross measures, not net. If nations with high Treasury/GDP ratios were also massive net borrowers, the Jones and Podemska (2010) result could be completely overturned. Fortunately, holdings of net foreign assets are also available, with annual data from 1970 to 2004 (Lane and Milesi-Ferretti 2004, 2007). This ratio of net foreign assets to GDP has the same positive predicted relationship with LV IQ. Omitting the OPEC countries, the phosphorous-rich micronation of Kiribati, and Liberia, the world's most FDI-intense nation, the correlation between LV IQ and net foreign assets to GDP in 2005 was +0.4; including all of these outliers increases the estimated regression coefficient, since Liberia is a massive recipient of foreign capital, though the *t*-statistic on IQ falls to 5. The relationship has strengthened over the period as barriers to capital flows have fallen since the end of Bretton Woods; in 1970 the correlation was +0.2.

Thus, mainstream growth theory combined with a conventional result in psychology (IQ's link to patience) can partially explain why some nations hold more financial and physical capital than others: Nations with higher average intelligence are more patient, and the patient inherit more of the earth.

Institutional Channels

Caplan and Miller (2010) find that within the USA, voters with higher IQs are more likely to support market-oriented policies, even controlling for income, education and political orientation. This might come as little surprise to those of us who have taught economics: The invisible hand is hard for some students to see, whether it comes to the unintended consequences of price controls or the power of the law of comparative advantage. Spontaneous order, multipliers, the law of unintended consequences: on average, the same individuals who on an IQ test can spot what is unusual in a drawing of a room full of children (one of the kids is facing the wrong way) are the same individuals who can see how doubling the

minimum wage is likely to hurt employment opportunities for the poor.

Since politicians tend to respond to the demands of citizens, whether wise or unwise, nations (especially democracies) with higher LV IQ are more likely to support the market liberalising policies that have been routinely found to be growth-promoting. And if individuals have a modest tendency to conform to their neighbour's views, the conventional finding of sociology, then the Caplan and Miller channel will become even stronger: people in slightly higher-IQ countries will have slightly higher-IQ neighbours, and tend to reinforce each other's slightly more liberal economic policy views. The urge to conform will create an IQ-voter quality multiplier.

O'Rourke and Sinnott (2006) provide preliminary cross-country evidence for the Caplan and Miller view: in the majority of countries, high-skilled workers are more supportive of trade than low-skilled workers. In the poorest countries, this relationship weakens, and it may reverse for countries at the GDP per capita level of the Philippines or below – nations with productivity 1/10th or less of frontier nations. Future research should investigate whether cognitive skills are good predictors of policy views in less developed countries.

Based on Caplan and Miller's results, one would predict that across the range of actually existing variation, high-IQ countries would tend to rank as freer on most indices of economic freedom. Unsurprisingly, this is the case: the Freedom House economic freedom measure correlates 0.6 with LV IQ (LV 2006, p. 251), and as we will see below national IQ also correlates positively with institutional quality.

Finally, smarter groups are more cooperative, more trusting, and more trustworthy in laboratory experiments (Pinker 2011, p. 611). If wealth-promoting political institutions depend partly on tacit cooperation among political elites, then this channel may help explain why high LV IQ nations are so much more prosperous. Jones (2008) was the first to find that high IQ groups were more cooperative in repeated prisoner's dilemmas: he collected data on repeated prisoner's dilemma experiments run at dozens of universities and found that when such experiments were run at high-SAT universities, students cooperated more often. This result held after controlling for whether the school was private or public, and a variety of experimental protocols such as number of rounds and whether students played for real money.

Later work has reinforced this result: Burks et al. (2009) in an experiment run on students in a truck driving school, found that in a one-round, two-move sequential prisoner's dilemma (similar to a Berg et al. (1995) trust game), high IQ players were more likely to cooperate in the first move, and were more likely to reciprocate cooperation in the second move. The first move corresponds to 'trust' and the second to 'trustworthiness'; high-IQ individuals possessed more of both traits in their sample.

Further, Putterman et al. (2010) found that among students at Brown playing a repeated public goods game, high-IQ players were more likely to contribute more in early rounds of the game, and contributed more overall. (Recall that the prisoner's dilemma is a two-action version of the public goods game.) And returning to Caplan and Miller's voter-quality channel, Putterman et al. found that when a voting round was added to the middle of the game, high IQ voters were more likely to vote for the efficient constitutional rule for punishing free riders. That IQ predicted pro-social behaviour at an Ivy League university reduces the likelihood that the results in other studies are driven by extreme social deprivation of low IQ individuals, anti-cooperation cultural norms among the families of low IQ individuals, or other similar sociological and environmental stories: among some of the world's most elite students, differences in IQ predicted differences in pro-social behaviour.

One unconventional measure (Jones and Nye 2011) finds that national average IQ and education levels are good predictors of law-abiding behaviour in a nearly lawless setting: the world of diplomatic parking in New York City. Until 2003, United Nations diplomats in New York were not required to pay parking tickets. Fisman and Miguel (2007) assembled this unpaid parking ticket data by country and found that diplomats from high-corruption countries were far more likely to earn parking tickets: corruption travelled

with the diplomat. But Jones and Nye find that when controlling for the national average IQ of the home country, the home country education level, or both, the statistical significance of home-country corruption is reduced or eliminated. The corruption channel may be operating through a human capital channel.

These results matter for political institutions, because politics is a repeated game where politicians are tempted to sacrifice long-run benefits for short-run benefits: high-level officials are tempted to take bribes that destroy transparency rather than reap the rewards of better institutions; judges and lawyers are tempted to collude rather than neutrally abide by the rule of law; members of parliament are tempted to confiscate capital after it has been invested in their nation (Jones 2011).

Repeated games abound in the public choice and political economy literatures, and in such models, a key parameter is always the discount factor, β. The discount factor is central to the folk theorem of repeated games: When players are more patient, they are better able to reach the Pareto-efficient solution. And as we have seen already, high-IQ individuals and groups appear to be more patient by a variety of measures. In a nation of patient players, politicians care more about long-run reputations (Persson and Tabellini 2000, Ch. 4), central banks find it easier to solve the time consistency problem (Barro and Gordon 1983) and thereby sustain a low-inflation equilibrium, and officials engaged in Rubenstein bargaining problems will split rents more equally, likely reducing social conflict.

Potrafke (forthcoming) finds cross-country evidence that national IQ is a reliable predictor of low national corruption as measured by the Corruption Perceptions Index, even when including a variety of historical and policy controls (Fig. 2). Surplus-destroying rent-seeking appears less common in high LV IQ countries.

If the links between IQ, patience, and pro-social behavior remain as strong as they appear in recent research, then differences in LV IQ are likely causing differences in the quality of institutions across countries. The political externalities of IQ may be large.

Maximizing Intelligence

What can be done to raise a nation's average level of intelligence? The public health literature has a set of obvious and data-driven answers: environmental improvements, childhood nutrition improvements, and better prenatal care all appear to be ways to increase IQ. A vast literature on the topic is summarized in Armor (2003); a few key pieces of evidence will need to suffice for the purposes of this essay. The link between environmental lead and intelligence is well-established; one recent paper, Ferrie et al. (2011) found that among Second World War draftees in the USA, higher exposure to lead through lead water pipes caused an IQ drop of five points. And in the Philippines, Solon et al. (2008) found that a 1 microgram increase in lead per litre of blood was associated with a 2.5 to 3.3 IQ point decrease in children. In the Filipino sample, children averaged 7.1 micrograms of lead per litre, a level that, extrapolated linearly, would predict at least a 15 IQ point decrease.

Until 2006, almost every sub-Saharan African country used leaded gasoline (United Nations Environment Programme 2002, 2005). The end of leaded gasoline in Africa will likely increase measured IQ in coming decades.

Experimental studies in developed and less-developed countries both suggest that for some individuals, IQ can be increased by providing proper micronutrients (Armor 2003; Jensen 1998, pp. 325–6). Nutritional channels are likely important in developing countries: in Pune City, India, a mere 10 weeks of zinc supplementation caused a 15–25 percent increase in student scores on the Ravens test (Tupe and Chiplonkar 2009). Notably, zinc supplementation increased the speed with which students pressed lighted buttons: Reaction time improved. Behrman et al. (2004) also survey evidence that increases in maternal and child health will increase IQ: iron and iodine deficiencies appear to be barriers to riches in the poorest nations.

A finding of the greatest importance is the Flynn Effect (1987), the well-known and conclusively documented trend of rising IQ across the rich countries. At least until the last decade, it appears that IQ scores have risen by two to three points per decade

across at least the second half of the twentieth century in the developed countries; while the debate continues over how much of this is nominal versus real (e.g. test-taking skill versus real-world problem solving and memory skill), there are sound reasons for believing that some portion is a genuine increase in mental ability. (One piece of evidence for a real increase in cognitive skill: human head size has increased by one standard deviation during the same time period when measured IQ in rich countries has increased by approximately the same amount (Jensen 1998, p. 325). If this is so, then economists should bring their unique tools to bear on the important question of why IQ has risen in the rich countries. Perhaps they will find ways to spur and strengthen a Flynn effect in the world's poorest countries. The most comprehensive discussion of the Flynn Effect is contained in a volume edited by Neisser (1998).

Finally, Eppig et al. (2010) have used the LV national average IQ data to argue that parasite prevalence, which correlates negatively with LV IQ, in fact helps to cause low national average intelligence.

If IQ provides some of the long-run positive externalities discussed in this essay – raising voter quality, improving institutions, providing access to frontier technologies, and raising capital intensity – then the benefits to improving public health are greater than currently believed.

Education may also increase overall intelligence; here the evidence is more scattered, but at the very least it appears plausible that increases in the quality and quantity of education raise measured IQ. Winship and Korenman (1997) drawing on the NLSY and comparing their results to other studies, conclude that one extra year of ostensibly exogenous education increased IQ by 'somewhere between 2 and 4 points' (p. 218). Hansen et al. (2004) used non-experimental methods to come to a similar conclusion. Card and Rothstein (2007) used plausibly exogenous variation in residency driven by desegregation court orders in the USA as in instrument for exogenous quality of schooling and peers. They found that an end to racial segregation in schools closes 'about one-quarter of the raw black-white gap in SAT scores', an IQ proxy (p. 2158).

The question remains whether these increases are what Jensen calls 'hollow IQ', mere test-taking skill rather than an increase in intelligence: perhaps future work can investigate whether the pattern of nerve conduction velocity and response times also moves in the expected direction when students exogenously receive increases in the quality and quantity of education.

One can also increase national average IQ quite reliably by allowing high-IQ individuals to immigrate. A decision to admit high-IQ immigrants as voting citizens would yield both the neoclassical and the institutional benefits of higher national average IQ. And since IQ correlates positively across generations, if high-IQ immigrants raise families in their destination country, then they will likely provide long-lasting benefits to the country that admits them.

If IQ has the sizeable positive externalities posited here, then there may be room for a Coasian bargain between countries with low current LV IQ and higher IQ individuals in other countries. One purely suggestive possibility: if the ratio of private to public benefits of higher IQ are even half as large as the 6:1 ratio suggested by Jones and Schneider (2010), a low LV IQ country could rationally offer a 100% subsidy for any wages a high-IQ immigrant earns in excess of that nation's median wage. In practical terms, a ten-year income tax holiday for permanent immigrants with engineering degrees could accomplish the same goal of encouraging high-IQ immigration.

A closing word on the question of possible genetically driven differences in national average IQ is needed. In developed countries, at the within-country, within-ethnicity level, it is clear that a substantial fraction of variations in IQ are genetically driven – around half or more (Caplan 2011; Boomsma et al. 2008; Plomin et al. 2000; Devlin et al. 1997). But the tools that behavioural geneticists use to establish the heritability of IQ or features such as height, eye colour or personality – twin studies and adoption studies – are rarely applicable for studying ethnic differences in intelligence, and of little use in studying cross-country differences.

Recent psychology textbooks and surveys (Loehlin 2000; Hunt 2010; Mackintosh 2011) discuss the issue of ethnic differences in

intelligence, partly drawing on adoption studies of children of East Asian and African descent adopted by families of European descent; here I discuss only the East Asian data. (Hunt's discussion of the James Watson affair (pp. 416–17) is recommended as a dispassionate summary of an important moment in the public discussion of the possible relationship between human genetics and intelligence.) One study cited by Loehlin and Mackintosh notes that adoptees of East Asian descent performed better than average on IQ tests; another repeated finding noted by these authors is that Native Americans, closely genetically related to East Asians, have a similar pattern of relatively high visual-spatial intelligence compared to whites. Together with other pieces of evidence, these findings induced both authors to either tentatively accept (Loehlin) or at least consider plausible (Mackintosh) the hypothesis that the higher average IQ of East Asians is at least partly genetic in origin. The hypothesis that at least some portion of cross-country IQ differences is genetic in origin has mainstream support from within the psychology profession.

But genetic differences do not imply intractable differences. The myopia common among high-IQ individuals is partly genetic in origin, yet eyeglasses and laser surgery have turned this genetic difference into a nuisance or less in wealthy countries. If IQ provides some of the long-run positive externalities discussed in this article – raising voter quality, improving institutions, providing access to frontier technologies and raising capital intensity – then the benefits to finding medical solutions for differences in IQ – whatever their origin – are greater than currently believed.

Conclusion

Since at least the work of Hanushek and Kimko (2000), economists have known that years of schooling are a poor measure of human capital; these authors found that national average test score measures were far better predictors of long-run economic performance. One weakness of the test score literature is that standardised mathematics, science and language scores are available for only a few dozen countries; another weakness is that little is known about non-school, non-wage individual-level correlates of such test scores (for expansions of these datasets, see Hanushek and Woessmann (2007, 2010)).

By using national average IQ as an index of human capital, growth economists can tap into a century of research by psychologists, sociologists, neuroscientists, geneticists and microeconomists into the causes, correlates and effects of IQ differences. Patience, pro-social behaviour and better-informed voting are among a few of the correlates of higher IQ discussed here, and others may exist. In cross-country growth regressions and calibrations, national IQ appears to have a much larger influence on economic outcomes than one would predict from conventional wage regressions; one can hope that future research will uncover practical methods for raising every nation's average level of cognitive skill.

See Also

- ▶ Behavioural Genetics
- ▶ Cognitive Ability
- ▶ Corruption and Economic Growth
- ▶ Easterlin Hypothesis
- ▶ Economic Growth
- ▶ Economic Growth, Empirical Regularities In
- ▶ Economic Growth in the Very Long Run
- ▶ Economic Growth Non-linearities
- ▶ Experimental Economics
- ▶ Growth and Institutions
- ▶ Human Capital
- ▶ Intertemporal Choice
- ▶ Level Accounting
- ▶ Time Preference

Acknowledgments I would like to thank Alex Tabarrok, Tyler Cowen and an anonymous reader for extremely helpful suggestions. Any remaining errors are my own.

Bibliography

Alderman, H., J.R. Behrman, D.R. Ross, and R. Sabot. 1996. The returns to endogenous human capital in Pakistan's rural wage labour market. *Oxford Bulletin of Economics and Statistics* 58(1): 29–55.

Armor, D.J. 2003. *Maximizing intelligence*. New Brunswick: Transaction Publishers.

Banerjee, A., and E. Duflo. 2011. More than 1 billion people are hungry in the world. *Foreign Policy*, May–June.

Barro, R.J., and D.B. Gordon. 1983. Rules, discretion and reputation in a model of monetary policy. *Journal of Monetary Economics* 12(1): 101–121.

Barro, R., and X. Sala-i-Martin. 2001. *Economic growth*. 2nd ed. Cambridge, MA: MIT Press.

Beaujean, A., M.W. Firmin, A.J. Knoop, J.D. Michonski, T.P. Berry, and R.E. Lowrie. 2006. Validation of the Frey and Detterman (2004) IQ prediction equations using the Reynolds intellectual assessment scales. *Personality and Individual Differences* 41(2): 353–357.

Behrman, J., H. Alderman, and J. Hoddinott 2004. Copenhagen consensus – Challenges and opportunities: Hunger and malnutrition. CopenhagenConsensus.com

Berg, J., J. Dickhaut, and K. McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10: 122–142.

Boomsma, D.I., T.C.E.M. van Beijsterveld, A.L. Beem, R.A. Hoekstra, T.J.C. Polderman, and M. Bartels. 2008. Intelligence and birth order in boys and girls. *Intelligence* 36(6): 630–634.

Bowles, S., H. Gintis, and M. Osborne. 2001. The determinants of earnings: skills, preferences, and schooling. *Journal of Economic Literature* 39: 1137–1176.

Burks, S., J. Carpenter, L. Goette, and A. Rustichini. 2009. Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences* 106: 7745–7750.

Caplan, B. 2011. *Selfish reasons to have more kids: Why being a great parent is more fun and less work than you think*. New York: Basic Books.

Caplan, B., and S.C. Miller. 2010. Intelligence makes people think like economists: Evidence from the general social survey. *Intelligence* 38(6): 636–647.

Card, D., and J. Rothstein. 2007. Racial segregation and the black–white test score gap. *Journal of Public Economics* 91(11–12): 2158–2184.

Deary, I. 2001. *Intelligence: A very short introduction*. New York: Oxford University Press.

Devlin, B., R. Daniels, and K. Roeder. 1997. The heritability of IQ. *Nature* 388: 468–471.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde. 2010. Are risk aversion and impatience related to cognitive ability? *American Economic Review* 100: 1238–1260.

Durbin, J. 1954. Errors in variables. *Review of the International Statistical Institute* 22(1/3): 23–32.

Eppig, C., C.L. Fincher, and R. Thornhill. 2010. Parasite prevalence and the worldwide distribution of cognitive ability. *Proceedings of the Royal Society B* 277(1701): 3801–3808.

Ferrie, J.P., K. Rolf, and W. Troesken 2011. Cognitive disparities, lead plumbing, and water chemisty: Intelligence test scores and exposure to water-borne lead among World War Two U.S. army enlistees. Working Paper, National Bureau of Economic Research.

Fisman, R., and E. Miguel. 2007. Corruption, norms, and legal enforcement: evidence from diplomatic parking tickets. *Journal of Political Economy* 115(6): 1020–1048.

Flynn, J.R. 1987. Massive gains in 14 nations: What IQ tests really measure. *Psychological Bulletin* 95: 29–51.

Frey, M.C., and D.K. Detterman. 2004. Scholastic assessment or *g*? The relationship between the Scholastic assessment test and general cognitive ability. *Psychological Science* 15: 373–378.

Hansen, K.T., J.J. Heckman, and K.J. Mullen. 2004. The effect of schooling and ability on achievement test scores. *Journal of Econometrics* 121(1–2): 39–98.

Hanushek, E., and D. Kimko. 2000. Schooling, labor force quality, and the growth of nations. *American Economic Review* 90: 1184–1208.

Hanushek, E., and L. Woessmann 2007. The role of school improvement in economic development. NBER Working Paper 12832, January.

Hanushek, E., and L. Woessmann 2010. The economics of international differences in educational achievement. NBER Working Paper No. 15949.

Hendricks, L. 2002. How important is human capital for economic development? Evidence from immigrant earnings. *American Economic Review* 92(1): 198–219.

Hunt, E.B. 2010. *Human intelligence*. Cambridge: Cambridge University Press.

Jensen, A.R. 1998. *The g-factor: The science of mental ability.* Westport: Praeger.

Jones, C. 2000. Comment on Rodriguez and Rodrick, 'Trade policy and economic growth: A skeptic's guide to the cross-national evidence. Manuscript, University of California.

Jones, G. 2008. Are smarter groups more cooperative? Evidence from prisoner's dilemma experiments, 1959–2003. *Journal of Economic Behavior and Organization* 68(3–4): 489–497.

Jones, G. 2010. The O-ring sector and the foolproof sector: An explanation for skill externalities. Working Paper, George Mason University.

Jones, G. 2011. National IQ and national productivity: The hive mind across Asia. *Asian Development Review* 28: 58–71.

Jones, G., and J.V.C. Nye. 2011. Human capital in the creation of social capital: Evidence from diplomatic parking tickets. Working Paper, George Mason University.

Jones, G., and M. Podemska. 2010. IQ in the utility function: Cognitive skills, time preference and cross-country differences in savings rates. Working Paper, George Mason University.

Jones, G., and W.J. Schneider. 2006. Intelligence, human capital and economic growth: A Bayesian averaging of classical estimates (BACE) approach. *Journal of Economic Growth* 11: 71–93.

Jones, G., and W.J. Schneider. 2010. IQ in the production function: Evidence from immigrant earnings. *Economic Inquiry* 48: 743–755.

Kremer, M. 1993. The O-ring theory of economic development. *Quarterly Journal of Economics* 108: 551–575.

Lane, P.R., and G.M. Milesi-Ferretti. 2004. The transfer problem revisited: Net foreign assets and real exchange rates. *Review of Economics and Statistics* 86(4): 841–857.
Lane, P.R., and G.M. Milesi-Ferretti. 2007. The external wealth of nations mark II: Revised and extended estimates of foreign assets and liabilities, 1970–2004. *Journal of International Economics* 73(2): 223–250.
Loehlin, J.C. 2000. Groups differences in intelligence. In *Handbook of intelligence*, ed. R.J. Sternberg. Cambridge: Cambridge University Press.
Lynn, R., and G. Meisenberg. 2010a. National IQs calculated and validated for 108 nations. *Intelligence* 38(4): 353–360.
Lynn, R., and G. Meisenberg. 2010b. The average IQ of sub-Saharan Africans: Comments on Wicherts, Dolan, and van der Maas. *Intelligence* 38(1): 21–29.
Lynn, R., and T. Vanhanen. 2002. *IQ and the wealth of nations*. Westport: Praeger.
Lynn, R., and T. Vanhanen. 2006. *IQ and global inequality.* Augusta: Washington Summit Publishers.
Mackintosh, N. 2011. *IQ and human intelligence*. Oxford: Oxford University Press.
Maurer, J., and A. Meier. 2008. Smooth it like the 'Joneses'? Estimating peer-group effects in intertemporal consumption choice. *Economic Journal* 118(527): 454–476.
Mischel, W., E.B. Ebbesen, and A.R. Zeiss. 1972. Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology* 21(2): 204–218.
Neal, D.A., and W.R. Johnson. 1996. The role of premarket factors in black–white wage differences. *Journal of Political Economy* 104(5): 869–895.
Neisser, U., eds. 1998. *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.
Neisser, U., G. Boodoo, T.J. Bouchard, A.W. Boykin, N. Brody, S.J. Ceci, D. Halpern, J.C. Loehlin, R. Perloff, R.J. Sternberg, and S. Urbina. 1996. Intelligence: Knowns and unknowns. *American Psychologist* 15: 77–101.
O'Rourke, K.H., and R. Sinnott. 2006. The determinants of individual attitudes towards immigration. *European Journal of Political Economy* 22(4): 838–861.
Peak, H., and E.G. Boring. 1926. The factor of speed in intelligence. *Journal of Experimental Psychology* 9(2): 71–94.
Persson, T., and G. Tabellini. 2000. *Political economics: Explaining economic policy.* Cambridge, MA: MIT Press.
Pinker, S. 2011. *The better angels of our nature: Why violence has declined*. New York: Viking.
Plomin, R., J. DeFries, G. McClearn, and P. McGuffin. 2000. *Behavioral genetics*. London: Worth Publishers.
Potrafke, N. (forthcoming). Intelligence and corruption. *Economics Letters*.
Putterman, L., J. Tyran, and K. Kamei. 2010. Public goods and voting on formal sanction schemes: An experiment. Working Paper, Brown University.
Quah, D.T. 1996. Empirics for economic growth and convergence. *European Economic Review* 40(6): 1353–1375.
Ram, R. 2007. IQ and economic growth: Further augmentation of Mankiw–Romer–Weil model. *Economics Letters* 94(1): 7–11.
Rindermann, H. 2007a. The big G-factor of national cognitive ability. *European Journal of Personality* 21: 767–787.
Rindermann, H. 2007b. The *g*-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality* 21: 667–706.
Rindermann, H., and J. Thompson. 2011. Cognitive capitalism: The effect of cognitive ability on wealth, as mediated through scientific achievement and economic freedom. *Psychological Science* 22: 754–763.
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(4): 813–835.
Shamosh, N., and R. Gray. 2008. Delay discounting and intelligence: A meta-analysis. *Intelligence* 36: 289–305.
Shoda, Y., W. Mischel, and P.K. Peake. 1990. Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology* 26(6): 978–986.
Solon, O., T.J. Riddell, S.A. Quimbo, E. Butrick, G.P. Aylward, M.L. Bacate, and J.W. Peabody. 2008. Associations between cognitive function, blood lead concentration, and nutrition among children in the Central Philippines. *Journal of Pediatrics* 152(2): 237–243.
Tupe, R., and S. Chiplonkar. 2009. Zinc supplementation improved cognitive performance and taste acuity in Indian adolescent girls. *Journal of the American College of Nutrition* 28(4): 388–396.
United Nations Environment Programme. 2002. Action plan for the phase out of leaded Gasoline in East Africa. http://www.unep.org/transport/pcfv/PDF/DataAPEAfrica.pdf
United Nations Environment Programme. 2005. Sub-Saharan Africa celebrates leaded petrol phase-out. Available online.
Volken, T. 2003. IQ and the wealth of nations. A critique of Richard Lynn and Tatu Vanhanen's recent book. *European Sociological Review* 19(4): 411–412.
Warner, J.T., and S. Pleeter. 2001. The personal discount rate: Evidence from military downsizing programs. *American Economic Review* 91(1): 33–53.
Weede, E., and S. Kampf. 2002. The impact of intelligence and institutional improvements on economic growth. *Kyklos* 55: 361–380.
Wicherts, J.M., C.V. Dolan, J.S. Carlson, and H.L.J. van der Maas. 2009. Raven's test performance of sub-Saharan Africans: Average performance, psychometric properties, and the Flynn effect. *Learning and Individual Differences* 20(3): 135–151.

Wicherts, J.M., C.V. Dolan, J.S. Carlson, and H.L.J. van der Maas. 2010a. A systematic literature review of the average IQ of sub-Saharan Africans. *Intelligence* 38(1): 1–20.

Wicherts, J.M., C.V. Dolan, J.S. Carlson, and H.L.J. van der Maas. 2010b. Another failure to replicate Lynn's estimate of the average IQ of sub-Saharan Africans. *Learning and Individual Differences* 20(3): 155–157.

Wickett, J.C., P.A. Vernon, and D.H. Lee. 2000. Relationships between factors of intelligence and brain volume. *Personality and Individual Differences* 29: 1095–1122.

Winship, C., and S. Korenman. 1997. Does staying in school make you smarter? The effect of education on IQ in the bell curve. In *Intelligence, genes and success: Scientists respond to the bell curve*. New York: Springer-Verlag.

Young, A. 2010. The African growth miracle. Unpublished manuscript, London School of Economics.

Zax, J.S., and D.I. Rees. 2002. IQ, academic performance, environment, and earnings. *Review of Economics and Statistics* 84(4): 600–616.

Ireland, Economics in

Tom Boylan and Renee Prendergast

Abstract

This contribution provides an overview of Irish economic thought from the 17th to the late 20 century. Broadly speaking, the pioneering contributions of the 17th and 18th centuries were concerned with the issue of improvement or economic development. The 19th century saw the formal institutionalization of political economy in Ireland and seminal contributions to value, distribution, and international trade theory in addition to work on public finance and methodology. The achievement of independence in the 20th century led to new concerns with development and policy experiments aimed at promoting lasting growth.

Keywords

Anglo–Irish Free Trade Agreement; Bastable, C.F.; Bentham, J.; Berkeley, Bishop G.; budget deficits; Busteed, J.; Butt, I.; Cairncross, A. K.; Cairnes, J. E.; Cantillon, R.; Carter, C.; Catch-up; Classical economists; Cliffe Leslie, T. E.; Common Agricultural Policy (EU); Concentration; Cost-of-production theory of value; Demand schedule; Development; Distribution theory; Dublin School; Duncan, G.; Dutch disease; European Economic Community; European Monetary Union; Factor immobility; Factor pricing; Fiduciary credit system; Foreign direct investment; Free trade; Geary, R. C.; Great Depression; Griffith, A.; Hancock, W. N.; Hearn, W. E.; Historical School; Hutcheson, F.; Imperfect competition; Induction; Infant-industry protection; Ingram, J. K.; Innovation; International trade; Ireland, economics in; Johnston, J.; Keynes, J. M.; Keynesianism; Kiernan, T. J.; Land question (Ireland); Lawson, J. A.; List, F.; Longfield, M.; Luxury; Lynch, P.; Malthus, T. R.; Mandeville, B.; Marginal revolution; Mill, J. S.; Mill–Bastable condition; Monetarism; National champions; Non-competing groups; O'Brien, G.; Oldham, C. H.; Owen, R.; Partnership; Paper money; Petty, W.; Planning; Protection; Rent; Research and development; Ricardian socialists; Ryan, L.; Self-sufficiency; Statistics and economics; Subjective theory of value; Swift, J.; Tariffs; Tax incentives; Technology policy; Terms of trade; Thompson, W.; Utilitarianism; Whately, R

JEL Classifications

B1

The 17th and 18th Centuries

William Petty and Richard Cantillon are commonly regarded as the founders of classical political economy. Both had connections with Ireland. Petty, English by birth, came to Ireland with the Cromwellian army in 1652 and became interested in 'political anatomy' in the course of surveying the country in preparation for the confiscation of Irish lands. Cantillon was born in Ireland but spent his adult life in as a banker in Paris, where he

wrote what many regard as the first systematic treatise on economics. Despite his nationality and his importance, Cantillon's work is not considered here. It was not written in Ireland; it was neither inspired by Irish conditions nor known to contemporaries living in Ireland.

Political Anatomy of Ireland, written in 1671–2, was Petty's first attempt to uncover the symmetry, fabric and proportion of the body politic by means of political arithmetic. Like all of Petty's writings, it contains pregnant theoretical suggestions, but our interest here is in its systematic approach to economic development. Petty sought to identify Ireland's development potential by considering the distribution and value of land and by estimating the number of 'spare hands' who could potentially add to local or universal (tradable) wealth. Petty identified the main causes of Irish underdevelopment as constraints on Ireland's trade with England and the plantations, insufficient coin, underdeveloped consumption patterns, perceived illegitimacy of rulers, rent-seeking and low population density. Petty's proposed remedies as set out in a *Report of the Council of Trade in Ireland,* 25 March 1676, included regularization of money, restoration of trade with the plantations and (particularly the cattle trade) with England, a bank based on landed property as security, reformation of the housing of the poor, legislative union with England and later the transportation of large numbers from Ireland to England. Despite a reference to discountenancing the use of certain foreign commodities in the report, Petty seems not to have favoured protection, arguing in *Political Anatomy* that the proceeds of exports would be more than sufficient to pay for imported products.

Partly as a result of prohibitions on the export of live cattle to England, farmers turned their attention to sheep, with the result that towards the end of the 17th century Irish wool and woollen yarn were among its most important exports. This promising development was nipped in the bud by restrictions introduced under the Wool Acts of 1698–9. This added to the fragility of an already weak economy, resulting in widespread poverty and unemployment in the early decades of the 18th century. Despite their confused and somewhat contradictory Irishness, the new generation of planter stock, including the likes of Prior, Dobbs, Browne, Molesworth, Hutcheson, Swift and Berkeley, responded with a steady stream of pamphlets advancing various proposals for improvement. These included increased agricultural investment, drainage and reclamation of bogs, improvement of inland waterways, encouragement of sea fisheries, mining and manufacturing, the setting up of a mint, consumption of locally produced goods, taxation of absentee rents, removal of restrictions on foreign trade and deportation of the undeserving poor (Kelly 1991). The main differences were between those such as Browne and Berkeley, who were relatively positive about Ireland's development prospects, and those such as Swift, who believed that plausible sources of improvement had little realistic chance of being implemented by those with the power to do so. Swift's position was vigorously expressed in *A Modest Proposal* (1729), a powerful satire on the pamphlet literature of his own time and one of the most telling critiques of positive economics ever to have been written anywhere.

While most authors emphasized the need to remove the constraints on trade, Berkeley argued that it would be more prudent to concentrate on those branches which were permitted, including Ireland's domestic trade (Berkeley 1752). Development would be possible even if the country were surrounded by a wall of brass. This, however, would require the substitution of domestically produced goods for the imported luxuries consumed by the elite as well as an expansion of the wants of the poorer classes in order to make them industrious. An argument for the reform of consumption patterns had already been made in 1726 by Francis Hutcheson in his 'Remarks upon the Fable of the Bees' in the course of controverting Mandeville's claim that luxury and vice were inseparable from economic development. Berkeley, who was also an implacable adversary of Mandeville, was even more emphatic than Hutcheson in his opposition to luxury, and he showed himself willing to contemplate sumptuary laws to achieve this objective. While Berkeley's proposals for development on the basis of the

domestic market were innovative at the time, his most radical proposals related to the adoption of paper money and the setting up of a national bank. Real wealth, Berkeley argued, consisted not in gold or silver but in the plenty of the necessaries and comforts of life and the power to command the industry of others. Money was simply a ticket or a counter for conveying or recording such power. As such, paper money and bank deposits were perfectly adequate and had some advantages over coin. The ruinous effects of the Mississippi and South Sea schemes were not due to paper money as such but to its use for speculative purposes rather than as a catalyst of industry. Private banks being subject to frauds and hazards, Berkeley proposed the setting up of a public bank, which he assumed would not suffer from these disabilities. The radicalism of Berkeley's position can be appreciated if we bear in mind that support for a fiduciary credit system as opposed to metallic money was in his time very much a minority view and remained so until recently (Murphy 2000).

The recovery of the Irish economy which took place in the second half of the 18th century was partly due to the success of the linen industry, which had been encouraged as a replacement for wool, and partly to the gradual weakening of commercial restrictions as Britain's population grew and Ireland became an important source of food and agricultural raw materials. During a brief period of legislative independence from 1782 to 1800, the Irish Parliament took steps to encourage domestic industry with various protective measures. It also introduced a corn law to encourage corn production for the British market.

The 19th Century

Following the Act of Union in 1801, Ireland was assimilated into the administrative and political jurisdiction of the United Kingdom. Many of its newly established industries went into gradual decline and corn production became a major source of employment. Population increased and with it poverty culminating in the Great Famine of 1845–50. These conditions influenced developments in political economy in Ireland and elsewhere. The need to counter the argument that high rents were a major cause of Irish poverty was a catalyst for the development of Malthus's rent theory (Prendergast 1987). The attention devoted by John Stuart Mill to the incentive effects of different forms of land tenure was partly a response to Irish land conditions. The scale of the human tragedy of the Irish famine influenced the perception and standing of laissez-faire political economy in Ireland and elsewhere. Irish economists became pioneers of the Historical School, which emphasized the specificity of time and place.

The formal institutionalization of political economy in Ireland began with the establishment of the Whately Chair in Trinity College in 1832. The chair was funded by Richard Whately, the Protestant Archbishop of Dublin, who came to Ireland from Oxford in 1831. The chair was part of a larger crusade by Whately to promote the dissemination of political economy with a view to encouraging more economically responsible behaviour. The Whately chair was to be filled by a number of outstanding occupants, which included Mountifort Longfield, John Elliot Cairnes and Charles Bastable. Chairs in jurisprudence and political economy were also established in the new Queen's colleges set up in at Belfast, Cork and Galway in 1845 (Boylan and Foley 1993). Outside of the universities, the principal institutional development was the founding in 1847 of the Dublin Statistical Society, later the Statistical and Social Inquiry Society of Ireland, which had Whately as its first president (Daly 1997). The society aimed at 'promoting the study of Statistical and Economical Science' and its participants included the academic, administrative and professional elite of Irish society. By the mid-19th century an extensive institutional infrastructure for the teaching and dissemination of political economy was in place (Boylan and Foley 1992).

Irish political economists in the 19th century made original contributions to a number of theoretical areas within the discipline. In value theory, the seminal contribution of Longfield, the first holder of the Whately Chair, has received

considerable attention and is recognized as providing one of the earliest attempts at formulating a subjective theory of value (Moss 1976). A number of Longfield's immediate successors, including Isaac Butt, James Anthony Lawson and William Neilson Hancock, also subscribed, albeit in a limited way, to a subjective theory of value, which led R.D.C. Black (1945) to suggest that the early Whately professors constituted a 'Dublin school' of subjective value theorists who anticipated by 30 years the marginal revolution of the 1870s. Longfield's contribution contained in his *Lectures on Political Economy* (1834) is by far the most original offering reflecting his disagreement with the dominant Ricardian framework of analysis in value and distribution theory.

Longfield approached value and distribution as pricing problems. The theory of value, in which commodity prices were determined in markets by supply and demand, was at the centre of his analysis. Longfield did not neglect the influence of cost on market price through changes in supply, but his main emphasis was on demand. The concept of a demand schedule was introduced, in which market demand was conceived as a ranking of individual demands according to their intensity, where 'the market price is measured by that demand, which being of the least intensity, yet leads to actual purchases'. Longfield invoked the concept of the individual's demand schedule as being composed of 'several demands of different degrees of intensity' (Longfield 1834, pp. 113, 114). This is now interpreted as a seminal statement, foreshadowing the principle of marginal utility that was to find its more formal articulation in the marginalist writers of the 1870s. Though not a member of the Dublin 'school', William Edward Hearn's *Plutology: Or the Theory of the Efforts to Satisfy Human Wants* (1863) was a significant contribution to the debate on the subjective theory of value. Hearn was appointed the first Professor of the Greek Language in Queen's College Galway in 1849. He left Galway in 1854 and became Australia's first professor of economics (Boylan and Foley 1984b). *Plutology* contained an extended and sophisticated taxonomy of the different kinds and degrees of human wants. Hearn went on to examine how demand could influence the impact of changes in the cost of production for different kinds of commodities; he distinguished between the demand for essential commodities or 'necessities' and non-essential or 'superfluities'. In this analysis Hearn provided a valuable extension to Longfield's earlier contribution, which was well regarded by contemporaries and later writers including Jevons and Marshall.

If Longfield and the Dublin 'school' represented an anti-Ricardian position in value and distribution, the Ricardian tradition was powerfully represented by Cairnes, the sixth holder of the Whately Chair at Trinity from 1856 to 1861 and Professor of Jurisprudence and Political Economy at Queen's College Galway from 1859 to 1870. Cairnes was arguably the most distinguished of the 19th-century Irish academic economists, and contributed to a number of areas of economic theory and contemporary policy issues. Cairnes was a close personal friend of J. S. Mill and was strongly influenced by Mill's analysis, but he produced a more complicated version of the theory of value than Mill. In *Some Leading Principles of Political Economy Newly Expounded* (1874), Cairnes provided a cost-of-production theory of value. But it is clear not only that Cairnes's 'normal value' is to be identified as cost of production, but that cost should be interpreted as real cost or sacrifice. In the course of his analysis he made the innovative move of applying Mill's proposition of the determination of international values by reciprocal demand in the case of factor immobility between countries to the internal economy of a country. In the latter situation, the existence of internal factor immobility gave rise to what is arguably Cairnes's most original application of the concept of non-competing groups. Cairnes's tenure in the Whately Chair broke the intellectual continuity of the Dublin 'school' by virtue of his commitment to the Ricardo–Mill approach.

In the domain of distribution theory one of the most interesting contributions was made by William Thompson (1775–1833), an Owenite and supporter of the French Revolution. Thompson pursued the aim of formulating an alternative economic system based on the rights of the primary

producer. He emerged as the most analytical and original thinker of the Owenite movement, which later became identified with the Ricardian socialists. Thompson was a personal friend of Bentham and it has been argued that Thompson's originality as a thinker consisted in his appropriation of the greatest happiness principle as a basis for fundamental social reform (Duddy 2002). While radical utilitarianism provided him with a critical component of his rationale for social reform, it was the adoption of Owen's system of mutual cooperation by Thompson, as a model of social organization, that would deliver to individual primary producers the fruits of their labour, which was fundamental to the Ricardian Socialists' doctrine.

In contrast to the Irish contributions to the Ricardian tradition of distribution theory, Longfield was forging a rather different approach in his *Lectures on Political Economy* of 1834. As Moss (1976) has argued, if the classical economists found the unifying principle for their theories of distribution in the concept of cost of production on the supply side, then Longfield could be said to have discovered his unifying principle of factor pricing in his supply and demand analysis. His identification of the role of marginal demand in the commodity market and marginal productivity in the factor market, justifies Longfield's claim as one of the leading progenitors of the neo-classical marginal theory of commodity and factor pricing.

In the area of international trade, the originality of Irish economists matched their contribution to value and distribution theory. In his *Three Lectures on Commerce and One on Absenteeism* (1835), Longfield extended the theory of comparative cost in significant directions, including the addition of both the multi-commodity and multi-factor case. He also addressed the issue of the incidence of tariffs and traced their effects on the relative price ratios between trading countries. Isaac Butt, Longfield's successor in the Whately Chair, considered the case for protection in his *Protection to Home Industry: Some Cases of Its Advantages Considered* (1846). This work, which was influenced by conditions in Ireland, was both methodologically engaging and analytically perceptive in its assessment of the benefits and weaknesses of protection in the context of particular circumstances.

Cairnes's reputation in the area of international trade rests on his systematic integration of the concept of non-competing groups into his analysis. This allowed him to distinguish between the role played by costs of production in determining international prices where effective competition existed; but where competition was absent, as in the case of non-competing groups, the fundamental determinant of international prices was not costs of production but reciprocal demand between noncompeting groups. He also provided an account of the factors determining the movements and range of a country's prices and money incomes arising from international trade, along with an original analysis of the process of international borrowing and the effects of loans on the equilibrium of international trade. This contribution that has been described as 'perhaps of greater permanent merit than any of his doctrines' in this area (Angell 1926, p. 94).

If the early and middle parts of the 19th century are associated respectively with the writings of Longfield and Cairnes, the latter part of the century must be identified with the work of Charles Bastable, who occupied the Whately Chair for 50 years, from 1882 to 1932. Bastable's *Public Finance,* first published in 1892, was a pioneering treatise that integrated, for the first time since McCulloch's *Taxation and the Funding System* (1845) what had become a rapidly expanding field of enquiry. Reviewing *Public Finance* in the *Economic Journal,* L.L. Price (1892) suggested it was the most comprehensive treatment of the topic since Adam Smith's *Wealth of Nations.* Bastable also made important contributions to international trade. In *The Theory of International Trade* (1887), he introduced varying elasticities of demand, increasing and decreasing returns and an extended analysis of obstacles to competition. In *The Commerce of Nations* (1892b), he provided a stringent critique of protection, while his name is associated with the celebrated 'Mill–Bastable' condition, which became an important part of the extended analysis of protection (Chipman 1965).

A distinguishing characteristic of many Irish economists in the 19th century was their commitment to an inductive method of approach. This was certainly true of the early Whately Professors. Isaac Butt maintained a robust scepticism with respect to the generality of economic principles, while Lawson was highly critical of Senior's efforts to reduce political economy to an axiomatic basis. It was not that the Irish writers called into question the validity of the deductive method in political economy. Rather, their position was that empirically observed facts should provide the basis for deductive reasoning. The methodological bias towards the inductive approach has been linked to the fact that the majority of the Irish professors were lawyers by training and profession and this allied to the preoccupation with the land question influenced their concentration on detailed studies of applied issues (Black 1947). Two of the most important representatives of the inductive approach were Thomas Edward Cliffe Leslie and John Kells Ingram. Both were major figures in the English-speaking world as pioneers of the Historical School of political economy. They were critics of the classical method of deduction and stressed the absolute necessity of an inductive approach to the study of economic issues, which in their view could never be separated from the larger social matrix of relations. The exception among the Irish contributions to economic methodology in the 19th century and to the inductivist position in particular was Cairnes who, in *The Character and Logical Method of Political Economy,* provided the most rigorous exposition of the deductive method that was produced in the course of the century.

The 20th Century

At the beginning of the 20th century, Belfast was Ireland's leading industrial centre, with strengths in linen, shipbuilding, rope making and engineering. Elsewhere agriculture predominated, and manufacturing was limited mainly to the food and drink industries. Against this background, the nascent independence movement regarded the development of industry as a matter of strategic importance. Drawing on the German economist Friedrich List, Arthur Griffith, the founder of Sinn Fein, proposed a programme for balanced economic development using protection on a broad scale. Protection was not to be permanent and was to be removed when the protected industries were strong enough to meet international competition (Griffith 2003). Industrial Development Associations throughout the country urged people to purchase Irish-made goods. Although the validity of the infant industry argument was widely acknowledged, in the main professional economists were not advocates of protection. Professor Oldham of University College Dublin argued that the relative openness and small size of the Irish economy meant that the protection of the home market could not provide a basis for development. Oldham was also concerned about the hidden costs of protection and suggested that bounties should be preferred to tariffs on grounds of their greater transparency and controllability (Oldham 1908, 1917).

During the 1921 Treaty negotiations with Britain, which were led by Arthur Griffith, professional economists including Riordan, O'Brien and Smiddy played a valuable role in securing the right of the Irish Free State to determine its future tariff regime (Girvin 1989). However, despite this and the fact that partition, which accompanied independence in 1922, involved the loss of Ireland's leading manufacturing centre, the new Free State government was cautious in its approach to economic policy and favoured free trade, fiscal prudence and the maintenance of the link with sterling. Bastable of Trinity and George O'Brien of University College Dublin were members of a committee set up in 1923 to consider the case for greater protection. The committee came out strongly against tariff protection for industry. Among the grounds given were that protection would raise costs for exporting industries, including the all-important agricultural sector, whose increasing efficiency and exports were seen as the main motor for growth.

The Great Depression of the late 1920s and the widespread protectionism to which it gave rise made a re-evaluation of the free-trade position necessary. In any event, a new government with

a different electoral base placed strong emphasis on self-sufficiency in both agriculture and industry. A bitter Anglo–Irish dispute over land annuities added further momentum to the protectionist drive. This led the Trinity College economist Joseph Johnston to argue in his polemical *Nemesis of Nationalism* (1934) that the Anglo–Irish dispute had been provoked by Eamon de Valera, the prime minister, in order to expedite his drive towards self-sufficiency. To judge from the contents of the Statistical and Social Inquiry Society journal during the period, the most prominent academic economists were also opposed to the policy of self-sufficiency. One of the few economists to comment favourably on the policy was J. M. Keynes, who also cautioned that only a modest degree of self-sufficiency could be achieved in such a small economy without a drastic impact on the standard of living (Whitaker 1983, p. 59).

The protectionist policies were successful in increasing industrial output and employment, but, as predicted by Keynes and as understood by Sean Lemass, the industry and commerce minister, and his top civil servant, the real challenge was to nurture industry to international competitiveness and to maintain the impetus for development once the initial easy phase of import substitution was over. In the event, the onset of the Second World War in 1939 and the growing scarcity of imported manufactures forced a further intensification of the policy of self-sufficiency. As elsewhere, the exigencies of the war economy led to greater government involvement in the allocation of resources and entrepreneurial activity generally. This continued after the war and, as late as 1959, Professor Charles Carter, formerly of Queen's University Belfast, commended the southern government for its willingness to engage in state enterprise if private enterprise failed to work, and contrasted this with the view taken in Northern Ireland that the function of government was to create the conditions for development and offer appropriate inducements but no more than that (Carter 1969).

In preparation for the aftermath of the war, policy debate on appropriate strategies for agriculture and employment took place in the early 1940s. A committee on agricultural policy chaired by T. A. Smiddy, De Valera's economic advisor, emphasized the importance of agricultural efficiency and the restoration of exports as a means of earning the foreign exchange that was necessary for the purchase of raw materials for industrial development. The other major policy debate of the period was occasioned by the publication in the UK of the Beveridge Report and the *White Paper on Employment.* The spectrum of Irish attitudes towards Keynesianism was reflected in a discussion of the problem of full employment held by the Statistical and Social Inquiry Society on 27 April 1945 (Lynch et al. 1945, 438–59). Opening the debate, Patrick Lynch, an economist in the Department of Finance and later Professor at University College Dublin, argued that the time had come to accept the Keynesian analysis of the economic system (Lynch et al., 438–41). The problems of the Irish economy were acute and required increasing state intervention. Government had to concern itself with the economy as a whole and not just its own expenditure as in the past. Lynch argued that government proposals for rural electrification and building were appropriate forms of intervention by means of which employment and further development could be stimulated. On the other hand, T. K. Whittaker, then number two at the Department of Finance, argued that Ireland was less exposed to cyclical fluctuations than Britain and America (Lynch et al., 446–9). Its unemployment problem was not primarily of a cyclical nature but the result of the insufficient investment in industry and agriculture. The problem was one of underinvestment rather than fluctuations in investment. Whitaker implied that increased investment in industry or agriculture would yield bigger returns at lower cost than the social investments mentioned by Lynch. Summarizing the debate, George O'Brien felt that there was general agreement that the Beveridge analysis did not apply to Irish circumstances (Lynch et al., 456–9). He noted that the main way in which Ireland had solved its unemployment problem during the last one hundred years was through the export of its people. Dr Beddy's recent paper comparing Irish and Danish agriculture (Beddy 1943–1944) had

shown that a more efficient agriculture was likely to employ fewer rather than more people. While the comparison with Denmark showed the possibilities offered by secondary industries, O'Brien himself felt that tertiary industries such as tourism had considerable potential.

While there were some attempts at policy and institutional innovation in the late 1940s and early 1950s, these involved attempts to make existing policy more effective rather than any major policy shifts. The performance of the economy was sluggish with low or sometimes negative rates of growth, and high levels of emigration. The publication of *Economic Development* (Department of Finance 1958) prepared by the Secretary of the Department, T. K. Whitaker, is commonly regarded as a major turning point in policy. The report demonstrates a remarkable consistency of position with the positions expressed by Whitaker in the 1945 debate on the *Beveridge Report.* The emphasis was on the need for productive investment. Whitaker argued that investment for which part if not all of the cost of servicing must be paid by the taxpayer was redistributive rather than productive, and should be replaced by productive investment. Despite the emphasis on the importance of investment, Whitaker cited A. K. Cairncross to the effect that entrepreneurial capacity was an even more important factor (1958, pp. 6–7), and in the body of the report he argued that the problem was not so much one of obtaining capital as securing 'know-what' as well as 'know-how' (1958, p. 154). These, it was suggested might have to come from external sources including foreign direct investment. The report also argued that, since the home market had been largely catered for, further development would have to depend on exports.

The 1958 Programme for Economic Expansion based on *Economic Development* involved a shift of emphasis from the promotion of domestically owned import-substituting industry to foreign-owned export-oriented industry. The implementation of the shift in policy over the following decade involved the easing of restrictions on foreign ownership and the implementation of incentives in the form of grants and tax exemptions on export generated profits. It also involved a shift from state-led enterprise to private enterprise, although Whitaker himself recognized that, since private investment was limited, productive investment would have to be engaged in by the public sector for some time to come. Another feature of the programme was its argument against existing policies of decentralization and in favour of the concentration of industries in large population centres with good internal and external communications and pools of skilled labour. A similar position on the concentration of new enterprises and infrastructure was later put forward in the Buchanan Report on regional development, which was published in 1968 and led to considerable debate (Buchanan and Partners 1968).

The introduction to the Programme for Economic Expansion emphasized that it was a not a plan and argued that the setting of detailed targets was inappropriate in a private enterprise economy exposed to fluctuations in external trade (Chubb and Lynch 1969). A few years later, there was much greater optimism about the value of planning. A second programme, which was developed in cooperation with the newly created Economic Research Institute (currently Economic and Social Research Institute, ESRI) and with input from Professor Louden Ryan of Trinity College, was much more specific in its targets. However, the actual performance of the economy in the period covered by the programme deviated from the planned targets, which were then abandoned. Drawing on this experience, the third programme emphasized the conditional nature of targets (Chubb and Lynch 1969). Despite this, the policies which were evolved in the programmes for economic expansion are widely regarded as providing the underpinning for the subsequent expansion of the economy. Governments also engaged in mildly expansionary fiscal policy, which helped to avoid the deflationary impact of Whittaker's own proposals. During the 1970s an attempt was made to counter the effects of oil price shocks by means of deficit spending. As the expected recovery failed to materialize, Ireland's debt burden grew and by the 1980s had reached unsustainable levels. During this period, Irish economists were vocal in their criticisms of fiscal policy and earned something of a reputation

as hard-nosed monetarists. The world recession of the early 1980s resulted in the closure of some foreign plants and made it difficult to attract new investment, so that there were net job losses in the foreign-owned sector. Irish economists criticized the Industrial Development Authority for subsidizing capital as a means of job creation. The Telesis review of industrial policy proposed a more selective approach to foreign direct investment, a shift of emphasis towards building strong indigenous national champions, the substitution of employment grants for capital grants and the setting up of a national linkage programme (NESC, 1982). The emergence of the economy from the 1980s recession, sometimes characterized as expansionary fiscal contraction, was achieved through a partnership agreement involving trade unions, employers and government. Partnership agreements have remained in place and are now regarded by economists in Ireland and elsewhere as an important industrial relations and development innovation (Teague and Donaghey 2004). In the decades since the 1960s, other policy issues with which economists engaged were the Anglo–Irish Free Trade Agreement, entry into European Economic Community, the Common Agricultural Policy and European Monetary Union. More recently, there has also been increasing emphasis on the factors governing the productivity and competitiveness of the economy as a whole.

When Ireland achieved independence in 1922, its cadre of professional economists was small, and the Statistical and Social Inquiry Society and its journal provided the main discussion forum for academic economists and government officials. Although some academics, including O'Brien, were literary in their approach, others such as George Duncan of Trinity College and John Busteed of University College Cork used a variety of statistical and empirical techniques. Duncan produced estimates of Irish national income to supplement T.J. Kiernan's pioneering efforts. Given that Duncan could be not regarded as anti-statistical, it is perhaps surprising that he was one of the main protagonists in a protracted debate between economists and Roy Geary. Geary was Ireland's foremost statistician but he also made important technical contributions to economics, including the Stone–Geary utility function as well as methods for updating input–output tables, for making international comparisons of real income, and for calculating the change in real income arising from changes in the terms of trade (Neary 1997; Spencer 1997). Geary argued that economics could become a science only through measurement and that economists' failure to appreciate the value of statistical work was due to their lack of awareness of the power of modern statistics. Geary also felt that economic theory was of very little value in the solution of practical problems and that academic economists were not sufficiently active in researching the social problems of the day. Duncan countered that the collection and manipulation of data could not by themselves advance knowledge of economic behaviour (Fanning 1984, pp. 151–5). He also pointed out that the Irish universities were seriously underfunded and had very little resources with which to carry out research. Part of the problem was a difference in attitude. Duncan, who had Austrian sympathies, disapproved of government intervention in general and of the protectionist policies of the day in particular. Geary, on the other hand, viewed policy issues as problems that could be solved with the correct technical means.

The present generation of Irish economists are more numerous and better trained than their predecessors in the early years of the 20th century. Many are the products of graduate schools in United States and in Britain. A recent examination of the journal output of Irish economists over the period 1970 to 2001 identified a total of 659 individual authors and 1,610 contributions, of which 218 were in the 1970s, 406 in the 1980s and 1,013 in the 1990s (Barrett and Lucey 2003). Of the total over the full period, half were in the main Irish journals: the *Economic and Social Review,* the *Journal of the Statistical and Social Inquiry Society of Ireland* and the *Irish Banking Review.* Other popular outlets for Irish economists were *Regional Studies Applied Economics* and the *Economic Journal*. A relatively small number of Irish economists contributed to the top international journals during the period surveyed by Barett and Lucey. Of these, Peter Neary made important contributions to the theory of international trade

as well as consumer theory, industrial organization and macroeconomics. In international trade, he is best known for work on Dutch Disease and the implications for trade of imperfect competition and technology policy.

In his seminal work on 20th century Ireland, the historian Joseph Lee (1989) argued that Irish economists have been impressive in the analysis of short-term movements but have contributed little to understanding the long-term development of the economy and have failed to contribute to development economics more widely. Whatever its truth in the past, this statement is no longer an accurate reflection of the state of affairs. Since the early 1990s, the Irish economy has experienced rapid growth so that its GDP per capita is now among the highest in Europe. This has led to considerable interest in understanding the nature and timing of the forces at work in Ireland's catch-up (Honohan and Walsh 2002). Meanwhile, however, Ireland's own development challenges have changed from those of catch-up to those of innovation and growth on the frontier. Meeting these challenges will require not only the strengthening of R&D capabilities but also addressing the special challenges of innovating in a small open economy.

See Also

- ▶ Agriculture and Economic Development
- ▶ Cairnes, John Elliott (1823–1875)
- ▶ Cantillon, Richard (1697–1734)
- ▶ Development Economics
- ▶ Foreign Direct Investment
- ▶ Geary, Robert Charles (1896–1983)
- ▶ Historical Economics, British
- ▶ Hutcheson, Francis (1694–1746)
- ▶ Ingram, John Kells (1823–1907)
- ▶ Petty, William (1623–1687)
- ▶ Swift, Jonathan (1667–1745)

Bibliography

Angell, J.W. 1926. *The theory of international prices: History, criticism and restatement*. Cambridge: Harvard University Press.

Barrett, A., and B. Lucey. 2003. An analysis of the journal article output of Irish- based economists, 1970–2001. *Economic and Social Review* 34(2): 109–143.

Bastable, C.F. 1887. *The theory of international trade*. 4th ed. London: Macmillan, 1903.

Bastable, C.F. 1892a. *Public finance*, 2003. Phoenix: Simon Publications.

Bastable, C.F. 1892b. *The commerce of nations*. London: Methuen.

Beddy, J.P. 1943–1944. A comparison of the principal economic features of Eire and Denmark. *Journal of the Statistical and Social Inquiry Society of Ireland* 17: 189–220.

Berkeley, G. 1752. *The querist.* Dublin. Reproduced in Johnston (1970).

Black, R.D.C. 1945. Trinity College Dublin, and the theory of value, 1832–1863. *Economia (n.s.)* 12: 140–148.

Black, R.D.C. 1947. Economic studies at Trinity College, Dublin – I. *Hermathena* 70: 67–80.

Boylan, T.A., and T.P. Foley. 1984a. John Elliot Cairnes, John Stuart Mill and Ireland: Some problems for political economy. In *Economists and the Irish Economy: From the eighteenth century to the present day*, ed. A.E. Murphy. Dublin: Irish Academic Press.

Boylan, T.A., and T.P. Foley. 1984b. Cairnes, Hearn and Bastable: The contribution of Queen's College, Galway to economic thought. In *Galway: Town and gown 1484–1984*, ed. D.O. Cearbhaill. Dublin: Gill & Macmillan.

Boylan, T.A., and T.P. Foley. 1992. *Political economy and Colonial Ireland*. London: Routledge.

Boylan, T.A., and T.P. Foley. 1993. The teaching of economics at the Queen's Colleges in Ireland (Belfast, Cork, Galway), 1845–1900. In *The market for political economy: The advent of economics in British University Culture, 1850–1905*, ed. A. Kadish and K. Tribe. London/New York: Routledge.

Boylan, T.A., and T.P. Foley. 2004. *John Elliot Cairnes: Collected works*. London/New York: Routledge.

Buchanan, C., and Partners. 1968. *Regional studies in Ireland*. Dublin: An Foras Forbartha.

Butt, I. 1846. *Protection to home industries: Some cases of its advantages considered*. Dublin/London: Hodges and Smith/John W. Parker.

Cairnes, J.E. 1874. *Some leading principles of political economy newly expounded*. London: Macmillan & Co..

Carter, C.F. 1969. Problem of economic development. In Chubb and Lynch.

Chipman, J.S. 1965. A survey of the theory of international trade: Part I, the classical theory. *Econometrica* 33: 477–519.

Chubb, B., and P. Lynch. 1969. *Economic development and planning*. Dublin: Institute of Public Administration.

Cliffe Leslie, T.E. 1888. *Essays in political economy.* 2nd ed. Dublin: Hodges, Figgis & Co..

Corden, W.M., and J.P. Neary. 1982. Booming sector and de-industrialization in a small open economy. *Economic Journal* 92: 825–848.

Daly, M.E. 1997. *The spirit of earnest inquiry: The statistical and social enquiry society of Ireland 1847–1997*. Dublin: Statistical and Social Inquiry Society of Ireland.
Department of Finance. 1958. *Economic development (The Whitaker Report)*. Dublin: Stationery Office.
Duddy, T. 2002. *A history of Irish thought*. London/New York: Routledge.
Fanning, R. 1984. Economists and governments: Ireland 1922–52. In *Economists and the Irish economy*, ed. A.E. Murphy. Dublin: Irish Academic Press.
Girvin, B. 1989. *Between two worlds – Politics and economy in independent Ireland*. Dublin: Gill and Macmillan.
Griffith, A. 2003. In *The resurrection of Hungary: A parallel for Ireland*, ed. P. Murray. Dublin: University College Dublin Press.
Hutcheson, F. 1726. Remarks upon the fable of the bees. *Dublin Journal,* 5, 12 and 19 February.
Honohan, P., and B. Walsh. 2002. Catching up with the leaders: The Irish hare. *Brookings Papers on Economic Activity* 2002(1): 1–77.
Johnston, J. 1934. *Nemesis of nationalism*. London: P.S. King.
Johnston, J. 1970. *Bishop Berkeley's querist in historical perspective*. Dundalk: Dun Dealgan Press.
Kelly, J. 1991. Jonathan Swift and the Irish economy in the 1720s. *Eighteenth- Century Ireland* 6: 7–36.
Lee, J.J. 1989. *Ireland 1912–1985 – Politics and society.* Cambridge: Cambridge University Press.
Leahy, D., and J.P. Neary. 1997. Public policy towards R&D in oligopolistic industries. *American Economic Review* 87: 642–662.
Longfield, M. 1834. *Lectures on political economy.* Dublin: William Curry, Jun. and Company.
Longfield, M. 1835. *Three lectures on commerce and one on absenteeism*. Dublin/London: William Curry, Jun. and Company/Longman and Company.
Lynch, P. et al. 1945. The problem of full employment: A discussion. *Journal of the Statistical and Social Inquiry Society of Ireland* 17, 98th session, 438–459.
Moss, L.S. 1976. *Mountifort Longfield: Ireland's first professor of political economy.* Ottawa: Green Hill Publishers.
Murphy, A.E. 2000. Canons of monetary orthodoxy and John Law. In *Contributions to political economy – Essays in Honour of R.D.C. Black*, ed. A.E. Murphy and R. Prendergast. London/New York: Routledge.
Neary, J.P. 1997. R. C. Geary's contributions to economic theory. In *Roy Geary, 1896–1983: Irish statistician*, ed. D. Conniffe. Dublin: Oak Tree Press.
Neary, J.P., and K.W.S. Roberts. 1980. Theory of household behaviour under rationing. *European Economic Review* 13: 25–42.
NESC (National Economic and Social Council). 1982. A review of industrial policy. A Report Prepared by the Telesis Consultancy Group. Dublin.
Oldham, C.H. 1908. The economics of 'industrial revival' in Ireland. *Journal of the Statistical and Social Inquiry Society of Ireland* 12: 175–189.
Oldham, C.H. 1917. Industrial Ireland under free trade. *Journal of the Statistical and Social Inquiry Society of Ireland* 13, part 96: 383–398.
Petty, W. 1672. Political anatomy of Ireland. In *The economic writings of William Petty*, ed. C.H. Hull, Vol. 1. Cambridge: Cambridge University Press 1899.
Prendergast, R. 1987. James Anderson's political economy – His influence on Smith and Malthus. *Scottish Journal of Political Economy* 34: 388–409.
Price, L.L. 1892. [Review of] Charles Bastable. *Public Finance Economic Journal* 2: 671–676.
Spencer, J.E. 1997. R. C. Geary: His life and work. In *Roy Geary, 1896–1983: Irish statistician*, ed. D. Conniffe. Dublin: Oak Tree Press.
Teague, P., and J. Donaghey. 2004. The Irish experiment in social partnership. In *The new structure of labor relations: Tripartism and decentralization*, ed. H.C. Katz and W. Lee. Ithaca: ILR Division, Cornell University Press.
Viner, J. 1955. *Studies in the theory of international trade*. London: George Allen & Unwin.
Whitaker, T.K. 1983. *Interests*. Dublin: Institute of Public Administration.

Irish Crisis: Origins and Resolution

John FitzGerald

Abstract

The Celtic Tiger years in the 1990s saw the standard of living in Ireland converge rapidly to the EU15 average. However, in the middle 2000s the rapid growth continued and demand rose well above potential output, driven by a property market bubble. Either appropriate fiscal policy or appropriate financial regulation could have prevented the ensuing crisis.

The management of the crisis after 2008 and the economic turnaround, which began in 2012, was reasonably successful. There was initially a period of very severe fiscal tightening, which brought the crisis in the public finances under control. Also, pre-emptive action by the government in building up liquid assets in 2008 and 2009 facilitated management of the crisis, though it did not obviate the need for extensive liquidity support through a bail-out programme.

While the tradable sector of the economy suffered from the boom and bust cycle it still survived reasonably intact. As a result, once the fiscal adjustment was completed, and with a return to world growth, the economy bounced back. The rapid recovery has been facilitated by the low interest rate environment and the fall in oil prices. While some of the lost ground as a result of the crisis will be made up, undoubtedly there will be a permanent loss of output as a result of the very severe crisis.

Keywords

Financial crisis; Financial regulation; Fiscal adjustment; Ireland

JEL Classifications

E32; E60; E62; E65

Introduction

In the early years of the 2000s, many foreigners sought answers as to how Ireland had turned out to be such a successful economy. An exceptional period of growth began in 1994, continuing into the 2000s. However, the bust of 2008 punctured such illusions – the behaviour of the Irish economy can be understood within a standard economic framework, albeit one that takes account of the unusual openness of that economy.

A key factor in the sustained period of economic growth was the past failure of the Irish economy. When Ireland became independent in 1922, it was substantially less well off than the UK. However, it was also better off than many other countries in Southern or Eastern Europe. In spite of this head start, its relative standard of living fell behind that of many other countries in Europe, especially in the years after the Second World War. Ó'Gráda (2002) argues that this underperformance owes much to domestic policy failures. It was only when these failures were addressed that the standard of living in Ireland converged to the EU15 average.

The major policy failings were the decision not to open up to free trade in the 1950s, the failure to invest in education in the immediate post-war years and unwise fiscal policy in the late 1970s. Exceptionally high tariff barriers were maintained against the outside world until the end of the 1950s, and it was only with EU membership in 1973 that the economy fully embraced the benefits and rigours of free trade. However, the way in which the Irish economy adapted to free trade depended, to an unusual extent, on foreign direct investment (Barry 2002).

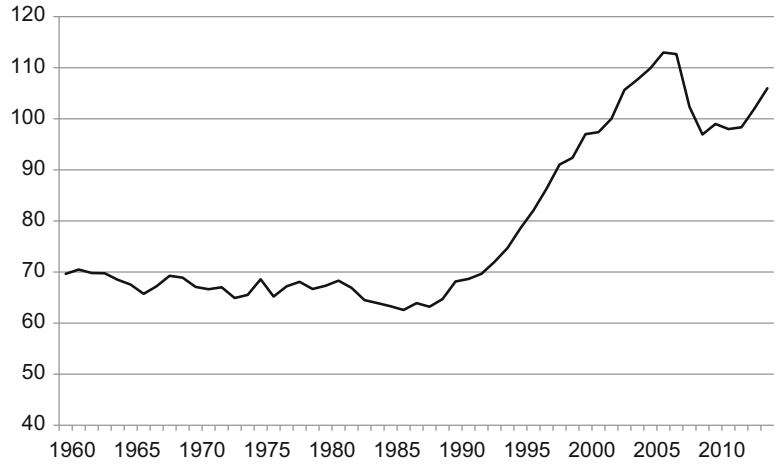
Free second level education was only introduced in 1967. At that time only around an eighth of children went on to third level education. This policy failure was gradually rectified over the next 20 years, but it was a slow process. It was only really in the 1990s that the effects of two decades of increasing investment in education began to impact on growth.

The convergence process, bringing the relative standard of living to where it should have been, was delayed by very unwise budgetary policies adopted at the end of the 1970s, which resulted in a fiscal crisis that took much of the 1980s to address. Thus it was only in the 1990s that the Irish economy converged to the EU15 average standard of living (Crafts 2014). When convergence happened, as shown in Fig. 1, the convergence process was very rapid.

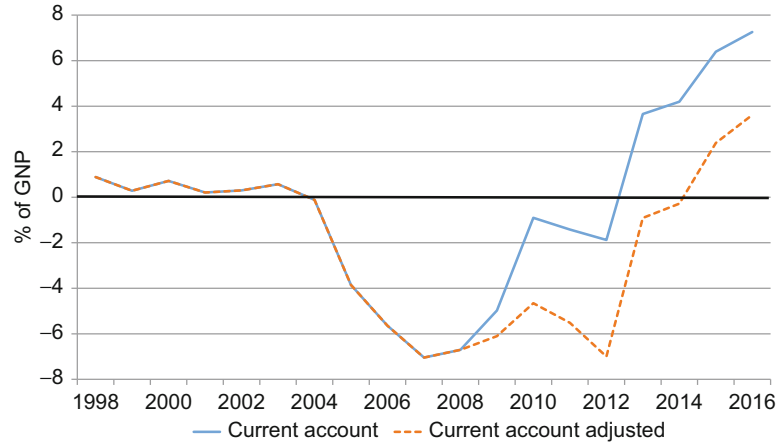
The next section examines the fit of hubris which resulted in a major real estate bubble in the middle of the last decade. We then consider the extent of the ensuing financial crisis and how the multiple problems were addressed through major policy changes. The penultimate section considers the nature of the economic recovery which began in 2012 and, finally, conclusions are drawn.

Hubris

The fact that the economy underwent a period of exceptional growth in the second half of the 1990s, when the increase in GNP averaged 9% a year, lulled policymakers into a false sense of security: after 2000 there was a growing feeling



Irish Crisis: Origins and Resolution, Fig. 1 GDP relative to EU15, adjusted for PPS, % (GNP for Ireland) (*Source*: Eurostat and CSO *National Income and Expenditure* for Irish GNP)



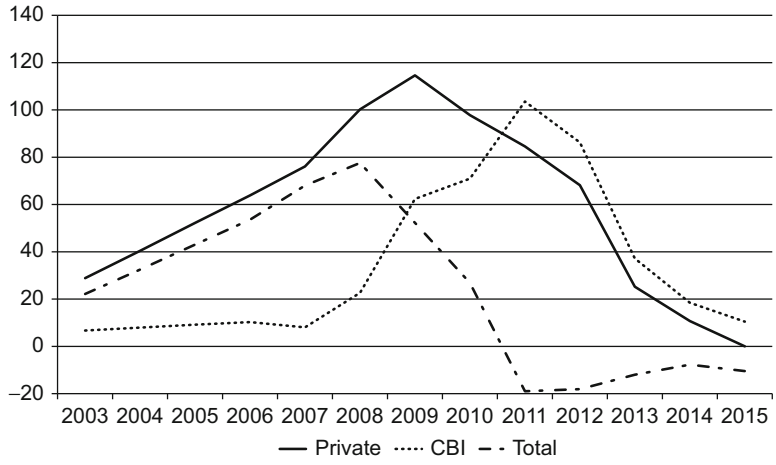
Irish Crisis: Origins and Resolution, Fig. 2 Current account of the balance of payments, % of GNP (*Source*: CSO Balance of Payments and additional data on inflow through redomiciled plcs. The adjusted figures exclude the revenue of these companies (FitzGerald 2015). Redomiciled plcs. are investment companies headquartered in Ireland, which don't generate real activity in the economy in terms of employment or purchases of domestic inputs)

that Ireland had found the elixir of eternal growth. (Throughout this article GNP is used rather than GDP because of the exceptional outflow of profits from multinationals, profits that are not part of national income.) As discussed earlier, the growth in the 1990s represented a catching-up process.

When, in 2001, the EU Commission suggested that the government needed to tighten fiscal policy to rein in the growth in demand, which appeared to be running ahead of potential output, this fell on deaf ears. The then government was very critical of this intervention by the EU Commission and it may have discouraged the EU Commission from criticising policy later in the decade. This policy advice, while not unreasonable, was probably premature: the bursting of the dotcom bubble in 2002 took some of the steam out of the economy, and the current account of the balance of payments remained in surplus till 2003 (Fig. 2).

Key factors in the rapid growth in the economy in the early years of the 2000s were foreign demand, which continued buoyant till 2007, and an investment boom. The economy needed to rapidly expand its infrastructure, private and

Irish Crisis: Origins and Resolution, Fig. 3 Investment as a share of GNP, % (*Source*: CSO National Income and Expenditure)



Irish Crisis: Origins and Resolution, Fig. 4 Net foreign liabilities of the banking system, % of GNP (*Source*: Central Bank of Ireland, Table A.4.1)

public, to cater for the growing population and the continuing increase in output. As shown in Fig. 3, investment, having run at between 25% and 30% of GNP up to 2003, peaked at over 35% of GNP in 2006. In the EU15 investment averaged around 20% of GDP over the last decade so the level of investment in Ireland was exceptional.

A significant factor in driving the investment share of GDP/GNP above 30% in the peak of the boom was the expansion of investment in housing. While Ireland had a greater need for investment in housing than many other EU15 countries because of the rapid growth in the population (Conefrey and FitzGerald 2010), housing investment accounted for an exceptional share of output in 2005 and 2006: around 14%, compared to the norm for other EU15 countries of around 4%.

Though domestic savings were enough to finance the high level of investment up to 2003, the further increase in investment thereafter generated a substantial and growing deficit on the current account of the balance of payments (Fig. 2), as domestic saving proved inadequate (Lane 2015a). Much of the investment in property was funded though the banking system. As shown in Fig. 4, the net foreign liabilities of the banking system rose from around 30% of GNP in 2003 to

almost 100% in 2008, peaking at over 110% in 2009.

The banks funded the expansion of their property lending by short-term borrowing on the interbank market. The resulting maturity mismatch posed massive problems for the banks when the crisis hit. The short-term funding dried up and they became hugely dependent on borrowing from the Central Bank (ECB), as shown in Fig. 4.

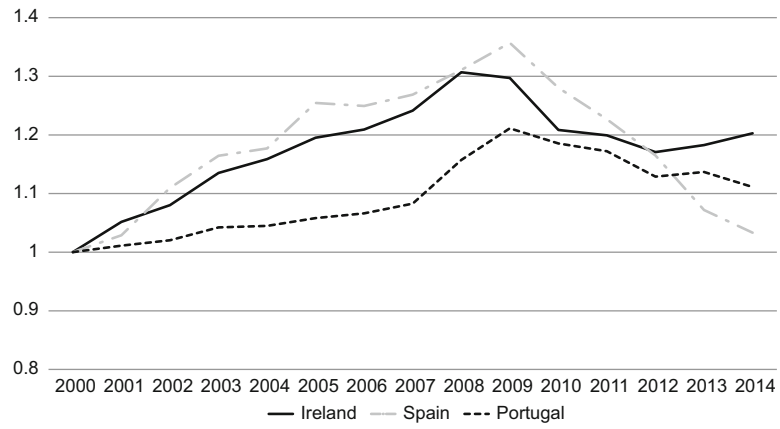
All of this pointed to the fact that the economy was growing beyond its potential. The labour market was also very tight, with large-scale immigration to meet the very rapid growth in employment. The unemployment rate was around 4.5%, full employment by Irish standards.

In the late 1990s the traditional outflow of emigrants from Ireland had been reversed and there was a substantial inflow of people coming to work in Ireland. They were generally well educated and played an important role in expanding the productive capacity of the economy (Barrett et al. 2002).

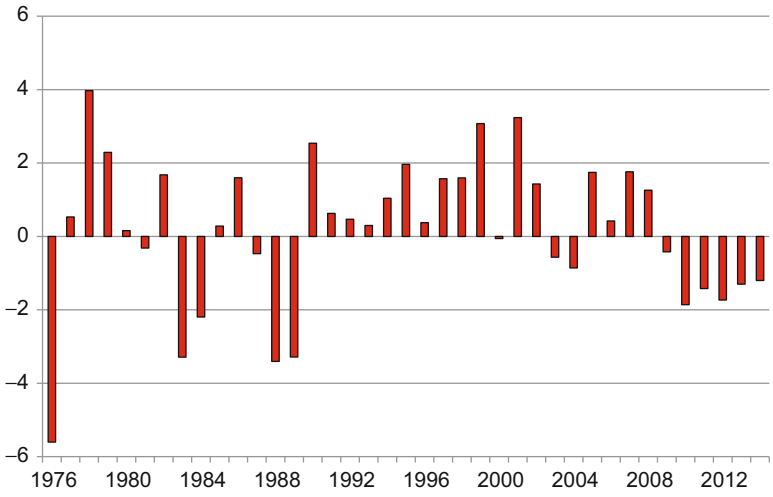
After EU enlargement in 2004 this inflow dramatically increased. Whereas before immigrants generally filled skilled jobs, a significant part of the post-2004 inflow moved into less skilled employment, including construction, possibly reflecting a lack of linguistic skills rather than a lack of education (Barrett and Bergin 2009). While the inflow of immigrants (and returning emigrants) helped relieve some of the labour market pressures, it also put upward pressure on the cost of accommodation.

The tightness in the labour market resulted in a rapid rise in wage rates relative to other EU15 countries (Fig. 5). The sectors related to building had to bid up wage rates to attract employees from other areas of the economy and from outside Ireland. This loss of competitiveness adversely affected the tradable sector of the economy: it was being crowded out by the growth in the size of the building and construction sector of the economy. In the run-up to the crisis, job losses occurred in the more labour-intensive parts of the tradable sector. The growth in the parts of the tradable sector employing predominantly skilled labour disguised the competitiveness problems faced elsewhere in the export sector. These less skilled, jobs lost in the tradable sector after 2005, have not reappeared, even with a reversal of the competitiveness loss.

While some domestic policy advice suggested that a significant tightening of fiscal policy was appropriate to prevent a property market bubble (FitzGerald et al. 2005; FitzGerald and Morgenroth 2006; Kelly 2007), this advice fell on deaf ears. As shown in Fig. 6, the stance of fiscal policy between 2000 and 2007 was generally stimulatory. In 2001 and 2002 the stimulus averaged over 2% of GNP a year. Again in 2005–2007 the stimulus averaged over 1.25% a year. (2002 and 2007 were election years and fiscal policy was generally stimulatory in the run-up to each election.) Much of this stimulus came from a rapid increase in public investment in infrastructure. However, this further aggravated



Irish Crisis: Origins and Resolution, Fig. 5 Average annual earnings relative to EU15 (*Source*: AMECO Database)



Irish Crisis: Origins and Resolution, Fig. 6 Fiscal stance, % of GDP. Positive is expansionary and negative is contractionary (*Source*: FitzGerald 2013a)

pressures in the building sector, given the concentration on investment on infrastructure projects.

Finally, as we now know (Honohan 2009, 2010; Oireachtas 2016) banking regulation was exceptionally lax. The result was a domestic banking system which was exceptionally exposed to a collapse in the property market in 2008: its assets were very illiquid, backed by very short-term borrowing.

Unlike Estonia, which saw a similar property-led bubble and collapse in 2007–2008, the bulk of the Irish banking system was domestically owned. While a few foreign banks (e.g. RBS) were also serious players in the market, making similar mistakes to the Irish banks, when the crisis hit the Irish government found itself responsible for the financial crisis in the Irish-owned banks. In the case of Estonia, because all of the banks were foreign-owned, the collapse in the economy did not result in the government having to assume a very large debt burden to bail out banks, though they still had to deal with the fiscal effects of the resulting economic crisis.

The Crisis

Over the course of 2007, house prices, which had risen rapidly for more than a decade, plateaued and began to slowly decline. Difficulties in the financial sector in the UK and the USA gave a warning of possible dangers for the Irish economy, but exports continued to grow rapidly. While some analysts published warnings concerning the housing market and the financial system (Barrett et al. 2007; Kelly, 2007) they were not heeded. With the benefit of hindsight, it was probably too late to avert a financial disaster, but earlier action would have reduced the subsequent damage.

The Budget for 2008 was published in December 2007 and it was based on an expectation of continuing growth, albeit at a lower rate than in the past. The dependence of government revenue, directly and indirectly, on the building and construction sector was not understood. The collapse in economic activity in the first half of 2008 began to have a dramatic effect on the fiscal position.

Banking Crisis

When Lehman Brothers was liquidated in mid-September 2008 it produced an even more concentrated focus on the potential problems of the Irish banking system. As a result, the Irish-owned banks faced a major funding crisis towards the end of the month. To deal with this the government provided a very extensive guarantee, covering most of the liabilities of the Irish-owned banks.

While the guarantee was supported in parliament in autumn 2008 by three of the four main political parties (Fianna Fáil, Fine Gael and Sinn Féin) the very wide scope of the guarantee has

been extensively criticised since it was introduced. It exposed the Irish government to huge liabilities, which crystallised in 2010 and 2011. Honohan (2009) argued that some form of guarantee was essential to prevent the collapse of the financial system, but that the guarantee as implemented was too wide.

The very wide criticism of this measure domestically must been seen against evidence that the ECB (and Ireland's partners within the EU) were not sympathetic to measures to impose losses on bondholders (Cardiff 2016). This restricted the government's freedom of movement, both in 2008 and again in the run-up to the bail-out in late 2010 and in the early months of 2011 (Oireachtas 2016). While the additional costs imposed on Ireland as a result of the refusal by the ECB to countenance the burning of bondholders probably amounted to at most 10% or 15% of the total cost of bailing out the banks, the ECB's action still rankles with the population.

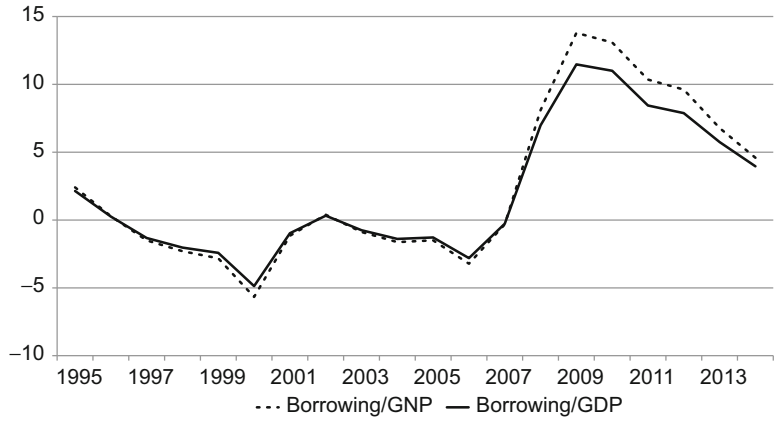
Fiscal Adjustment

The Budget for 2009 was passed early in October 2008. The major parameters of the budget had already been determined before the banking crisis of late September, and the forecasts for 2009 and the resulting estimates of revenue and expenditure proved totally unrealistic. This became apparent well before the end of the year. In spite of additional budgetary measures early in 2009, government borrowing, which had reached 11.5% of GNP in 2008, peaked at almost 14% of GNP in 2009 (Fig. 7). Because of the unexpected fall in inflation, the budget for 2009, even after amendment, was only mildly contractionary (Fig. 6). The real value of welfare payments actually rose significantly in 2009 because of the fall in prices.

Well before the gravity of the crisis became apparent to the wider public in late 2008, the National Treasury Management Agency (NTMA) began a major drive to fund the government. Because the gravity of the situation was not fully apparent, they were able to raise a large amount of money at reasonably attractive interest rates in 2008. They continued this funding drive in 2009, converting short-term borrowing into bonds with longer maturities.

As a result, by the end of 2008 government holdings of cash amounted to 16% of GDP, increasing to 18% of GDP by the end of 2009, in spite of the huge deficit in both years. This move to provide a large liquidity buffer contrasted with the positions in Portugal and Greece, other troubled economies. In the case of Portugal, the move to provide a large liquidity buffer only began in 2011.

While there was an understanding that problems in the banks would require a significant injection of capital by the government, the full magnitude of the problem was not understood in 2009. As a result, the government felt that the liquidity buffer, which amounted to the expected funding needs for 2009 and 2010, would be sufficient for Ireland to ride out the financial storm based on its own resources. However, the problems in the banking sector proved to be much greater than anticipated, and the crisis in the EU



Irish Crisis: Origins and Resolution, Fig. 7 Government borrowing, % of GDP and GNP (*Source*: CSO National Income and Expenditure and Government Financial Statistics)

financial markets, which began with the problems in Greece, meant that even the large cash buffer available at the beginning of 2010 was not adequate. The ECB's refusal to countenance imposing costs on senior bondholders in the banks added to the government's financial pressures in 2010–2011.

In 2009 the government had set up the National Asset Management Agency (NAMA) to buy bad property loans from the banks at a suitably discounted price. Initially in 2009 there were fears that NAMA would overpay for the loans, providing a subsidy to the troubled banks. However, when NAMA took on the first major tranche of loans early in 2010 it applied a fairly severe discount. This made it apparent to the government that a major cash injection into the banks would be needed. However, when NAMA took on the second major tranche of loans in early summer 2010, and applied an even bigger discount, it became apparent to the financial markets that the banks would need a very large government injection, an injection which made even the large cash buffer that the government held look inadequate.

The result was a loss of confidence in the financial markets in the autumn of 2010, which eventually forced the government to seek assistance from its EU partners and the IMF in November 2010. Before this assistance was sought the government published a medium-term adjustment plan which promised a very severe Budget for 2011, to be introduced in December, together with further major adjustment in the following three years. The planned borrowing for 2011–2013 is shown in Table 1. The 'Troika' of the EU Commission, the ECB and the IMF accepted the fiscal adjustment plan, which was already in place, without seeking significant changes. Thus the adjustment programme was not 'imposed' by the Troika, but decided by the Irish government.

In an election at the beginning of 2011 the outgoing government suffered massive losses and was replaced by a new administration. However, the incoming government adopted the broad parameters of the adjustment plan already in place, while making significant changes in the detailed measures.

Irish Crisis: Origins and Resolution, Table 1 Planned government borrowing, % of GDP

	2010	2011	2012	2013
Plan of: **Spain**				
Spring 2010	9.8	7.5	5.3	3
Spring 2011	9.2	6	4.4	3
Spring 2012			5.3	3
Latest	9.4	9.4	10.3	6.8
Plan of: **Ireland**				
Winter 2009	11.6	10	7.2	4.9
Winter 2010		10.6	8.6	7.5
Latest	11.0	8.4	7.9	5.8

Source: Stability programme updates for Spain and Ireland and EU AMECO database

Irish Crisis: Origins and Resolution, Table 2 *Ex ante* fiscal adjustment, % of GNP

	2009	2010	2011	2012	2013	Total
Revenue	3.9	0.0	1.0	1.1	0.7	6.8
Expenditure	2.8	3.1	2.8	1.5	1.5	11.6
Total	6.7	3.1	3.8	2.7	2.2	18.4

Source: Department of Finance Budgets and CSO: *National Income and Expenditure*

Table 2 shows the headline increases in taxation and cuts in expenditure in each budget from 2009 to 2013 (FitzGerald 2013a). (*Ex post* the outcome is different, as the effect of the fiscal tightening reduced activity and, hence, reduced tax revenue and increases expenditure.) This shows the 'fiscal effort', but it takes no account of inflation. In particular, what looked like a very severe budget in 2009 turned out to be much less severe (Fig. 6), due to the fall in the price level, which turned nominal cuts in welfare into real increases. The cumulative adjustment was very large, with about 40% of the effort coming from increases in revenue (taxes) and the rest from cuts in expenditure, including a large cut in public service pay rates.

A feature of the adjustment plan put in place by the outgoing government in November 2010 was that it 'under-promised'. The plan took a deliberately conservative view on the public finances in spite of the fact that the government faced an imminent election. When the incoming government took power it was able to adopt the plan knowing that the targets set in it, while very

tough, would be achievable. As a result, the new government was able to exceed the agreed fiscal targets every quarter for the next three years. This helped restore confidence in the Irish economy among financial markets and, eventually, among citizens.

Irish Adjustment in an EU Context

As Table 1 shows, while the outgoing Irish government under-promised, facilitating the incoming government over-delivering, the position was different in Spain. There the outgoing government set unrealistic adjustment targets in 2011, which meant that the incoming government appeared to fail against this unrealistic benchmark. This made the adjustment in Spain look less successful than in Ireland in the early years of the new Spanish government, in spite of the substantial fiscal tightening that was actually taking place.

The crisis in Ireland, which was reflected in an unsustainable current account deficit, resulted from excessive investment, in particular in housing and commercial property, not from excessive private or public consumption. This investment bubble seriously damaged the rest of the economy: the inflated building and construction sector sucked resources from the rest of the economy through raising prices, especially the price of labour. With the benefit of hindsight, while the collapse of the property market that occurred in 2008 was triggered by the global financial crisis, the price level and the size of the building sector were not sustainable (Whelan 2014).

When the bubble burst the investment demand collapsed, driven by market forces, and this produced a dramatic turnaround in the current account deficit, helping put the economy on a sustainable footing. This reduction in demand was not caused by fiscal policy, although it had massive consequences for the public finances. Thus it happened 'automatically' without a need for approval by parliament.

The pattern was rather similar in Spain, Estonia and Latvia: an unsustainable investment boom which collapsed, helping to restore balance on the external account (FitzGerald 2013b). In these countries the public finances were in surplus in the boom years, but developed very large deficits with the collapse in investment. After the event it was clear to most people that the investment bubble had been unsustainable and that governments could not protect those directly affected working in the building-related sectors.

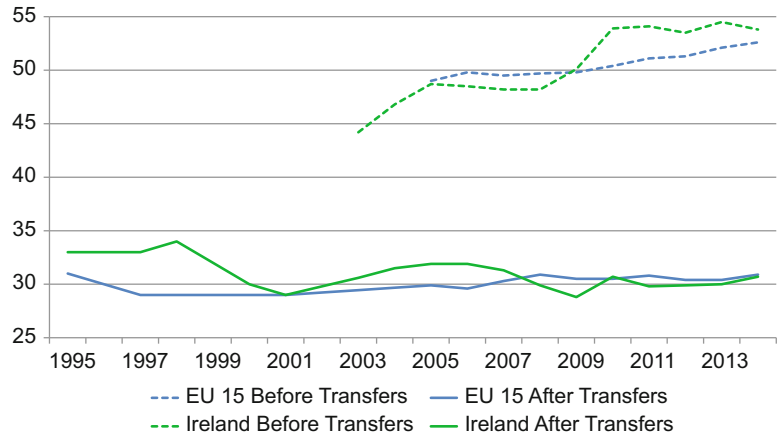
For Greece the situation was rather different. There the imbalance on the external account arose from an excess of consumption: public and private. Investment as a share of GDP was not excessive. Addressing this problem required major fiscal action and, hence, parliamentary approval. However, it was much more difficult politically to undertake such an adjustment, which was going to directly affect the living standards of the whole population. While a collapsing bubble happens very rapidly, producing a rapid adjustment in the external imbalance, undertaking a major fiscal adjustment programme takes a number of years.

Distributional Effects

From its peak in 2007 to the trough early in 2012, employment fell by almost 16%. The fall in employment was particularly severe in the building and related sectors. In spite of very substantial net emigration, especially by those who had lost their jobs, the unemployment rate rose from 4.6% in 2007 to over 15% of the labour force in 2011. The families of those who lost the jobs were the biggest losers from the crisis.

While the unemployment rate rose from under 5% of the labour force before the crash to peak at 15% in 2011, it could have been much worse. Employment fell by 16% points from peak to trough and there was a continuing large net inflow into the working age groups. However, as had happened on many occasions over the previous century, when the Irish labour market suffered a major downturn many of those who lost their jobs sought employment elsewhere. This applied to both the foreign workers who lost their jobs in the construction sector and to Irish-born of working age. This pattern of immigration and emigration, depending on prevailing labour market circumstances, reflects the extreme openness of the Irish economy. It has been a major shock absorber for the economy over the last century.

At the same time as unemployment increased, many of those on very high incomes, earned from



Irish Crisis: Origins and Resolution, Fig. 8 Gini coefficient before and after government transfers and taxes (*Source*: Eurostat)

Irish Crisis: Origins and Resolution, Table 3 Government transfers as% of GDP (For Ireland GNP)

								Change
	2007	2008	2009	2010	2011	2012	2013	2007–2011
Germany	16.0	15.8	17.4	16.7	15.7	15.6	15.7	−0.3
France	17.4	17.6	19.2	19.2	19.1	19.5	19.9	1.7
Netherlands	9.7	9.7	10.7	11.0	11.1	11.5	11.9	1.4
UK	12.1	12.6	14.3	14.2	14.2	14.6	14.5	2.1
Ireland	**11.5**	**13.8**	**17.7**	**17.6**	**17.5**	**17.5**	**16.3**	**6.0**
Greece	14.6	16.1	17.6	17.8	19.3	19.8	18.5	4.7
Spain	11.5	12.3	14.4	15.1	15.3	16.0	16.3	3.8
Portugal	14.1	14.6	16.4	16.4	17.0	17.5	18.4	2.9

Source: EU AMECO Database

involvement in the property boom, also lost heavily. However, the losses at the top of the income distribution did not compensate for the losses at the bottom of the distribution, so that the distribution of market income, which was already quite unequal by EU standards, became even more unequal. Figure 8 shows the Gini coefficient, based on market income, before government transfers, for Ireland and the EU15. (The higher the coefficient the more unequal the distribution of income.)

Figure 8 also shows the Gini coefficient after transfers and taxes for both Ireland and the EU15. This illustrates how, as a result of transfers, inequality fell slightly in Ireland from the boom years of 2006–2007 to 2013. The widening gap between the Gini coefficient for Ireland on a market incomes basis and on an after tax and welfare basis reflects a decision made by successive governments to protect welfare recipients, only making a moderate reduction in welfare rates in nominal terms.

As shown in Table 3, the share of transfers in GDP rose by 6% points over the course of the crisis, putting huge pressures on the budget. With a collapse in revenue, this increase in the welfare budget necessitated even more severe increases in tax rates and cuts in other areas of expenditure. In Greece, Spain and Portugal, which underwent a similar crisis, the increased share of transfers in national income was significantly lower than in the case of Ireland.

The Recovery

While the fiscal adjustment was still having a negative effect on the economy in 2012 and into 2013, the tradable sector had continued to grow through the very deep recession. With growth in

the world economy, especially in the USA and the UK, and a restoration of competitiveness, the economy turned the corner in 2012, showing a return to growth in GNP in the latter part of that year. By 2013 the economy was growing rapidly, and that rapid growth continued into 2016.

Initially the growth was driven by the external sector. However, there was a significant increase in investment in 2014 (Fig. 3) from a very low base and, in 2015, there was eventually a return to growth in consumption. Because of the fact that the recovery has been led by the growth in external demand for Irish goods and services, the current account, which was still in deficit in 2012, had moved back into balance in 2014 (Fig. 2), suggesting that the recovery is sustainable.

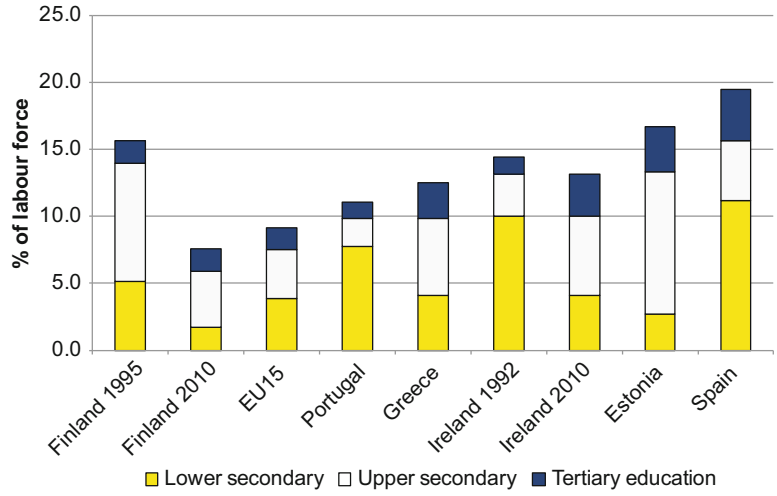
During the recession years it was feared that the dramatic rise in the unemployment rate would be very difficult to reverse. Experience in other countries, and also the experience of the exit from the 1980s crisis in Ireland, suggested that the recession might have a permanent effect on the labour market. People out of work for some time may lose skills and may find re-entry into the labour market difficult.

As shown in Fig. 9, the educational attainment of those who were unemployed in Ireland as the crisis was at its worst in 2010 was very different from when Ireland was exiting from the previous crisis in 1992. In 1992 over two-thirds of those who were unemployed had at most lower secondary education. In 2010 the situation was reversed, with over two-thirds having at least upper secondary education. This contrasts with the situation in Spain and Portugal, where most of those out of work in 2010 had very limited education.

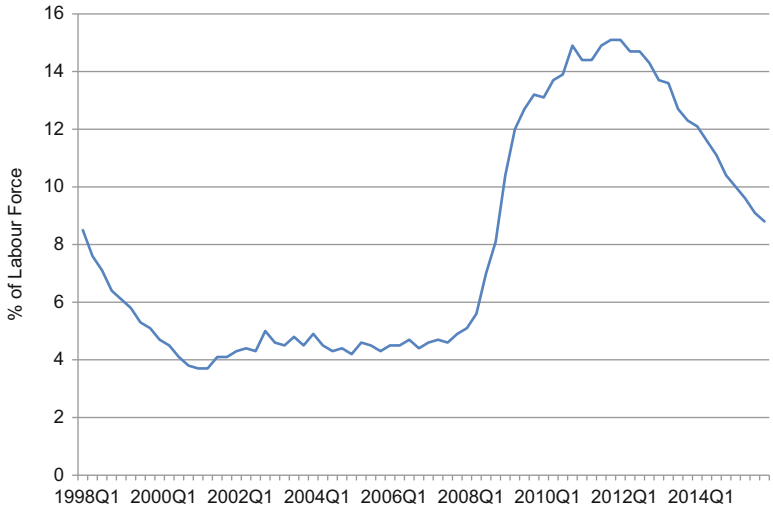
For those with at least an upper secondary education the prospect of finding work in a recovering Irish economy was good. Also, past experience had shown that, for those with a good education, where suitable jobs are not found in Ireland, they tend to emigrate to find work in other buoyant labour markets.

Since 2012 the rate of unemployment has fallen continuously in Ireland (Fig. 10). Also the rate of long-term unemployment has declined in nearly every quarter, indicating that a feared hysteresis effect on unemployment was exaggerated. The figures certainly suggest a smoother return to work after the recent crisis than was the case in the 1990s.

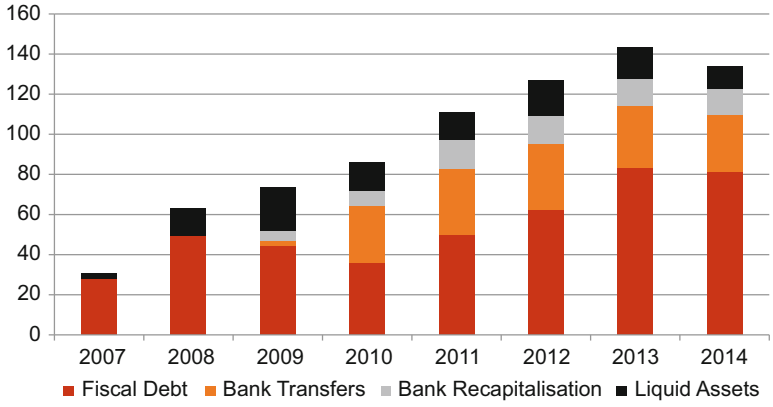
Another legacy of the crisis, which impacted on the recovery, was the very high level of indebtedness in the household sector. Having borrowed heavily to fund the building boom of the last decade, many households found themselves in 2012 heavily indebted and in negative equity, in spite of some growth in property prices. However, there is a cohort effect operating here: for those under 35 the level of indebtedness is low, with many households having some savings (Byrne



Irish Crisis: Origins and Resolution,
Fig. 9 Unemployment by level of education (*Source*: Eurostat)



Irish Crisis: Origins and Resolution,
Fig. 10 Unemployment rate, % of the labour force (*Source*: CSO Quarterly National Household Survey)



Irish Crisis: Origins and Resolution,
Fig. 11 Composition of Irish gross debt, % of GNP (*Source*: CSO: National Income and Expenditure, Government Finance Statistics; NTMA Annual Reports)

et al. 2014). This is because they were either too young or too wise to buy during the boom years. It is the 35–45 cohort who are particularly indebted and constrained in their consumption behaviour.

Initially this high level of indebtedness prevented a return to consumption growth. In addition, with falling prices there was little incentive for households to buy new dwellings in the period to 2012. However, growing demographic pressures, combined with a return to growth in house prices, have resulted in serious pressure on the housing market. To date the supply response has been very limited, with housing investment well below the 'normal' level of the last 40 years. This is reflected in rapidly rising rents.

If and when housing supply responds, this will add to the growth in domestic demand. The lessons from the past suggest that such a recovery will need careful management. While demographic pressures require at least a doubling in housing output, it is not clear that domestic savings will be adequate to fund such a recovery. The Central Bank has introduced macroprudential measures to prevent households and banks making the same mistakes as they did in the last decade.

At the height of the crisis there were major fears that the level of government debt would be unsustainable. Figure 11 shows the origins of the government debt as a percentage of GNP. As discussed earlier, and illustrated in the figure, there is a significant difference between the gross debt and the debt net of government cash holdings.

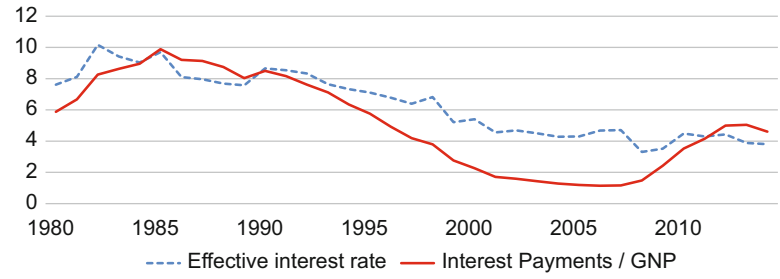
The injection of capital into the banks at its peak in 2011 accounted for 47% of GNP. However, between 2007 and 2013 the bulk of the increase in indebtedness was accounted for by the accumulated borrowing to fund the huge deficit (55% of GNP).

Since 2013 the burden of the debt has begun to fall, with growth in the economy and some initial repayments from the banking system. In the long run, if the recovery continues and is successfully managed, the state could benefit from the eventual sell off of the state-owned banks, recovering some, but not all of its forced investment in the banks.

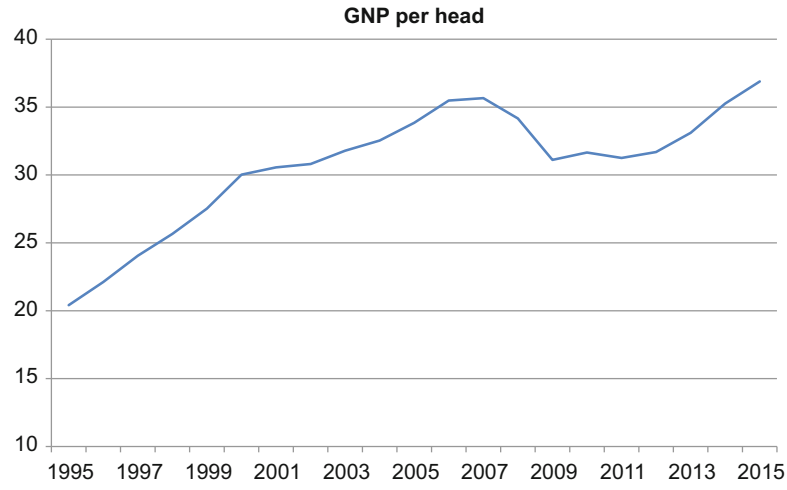
While the debt to GNP ratio is extremely high, the low rate of interest on new borrowing by the government has allowed it to refinance much of the debt on favourable terms. As shown in Fig. 12, interest payments accounted for just over 4% of GNP in 2014. In the previous crisis they peaked at 10% of GNP. Also, the average interest rate on the debt was around 4% in 2014, and this is likely to fall as the existing debt is refinanced at current low bond yields.

Conclusions

Over the last decade the Irish economy has experienced an exceptionally deep recession and a financial collapse, imposing a huge burden on the state because of the need to recapitalise the domestically owned banks. As shown in Fig. 13, GNP per head fell by 13% between 2007 and 2009. This fall in output was concentrated in the non-tradable sector of the economy, especially those sub-sectors related to building and construction. The tradable sector of the economy came through relatively unscathed. With a return to competitiveness and a return to growth in external markets, the tradable sector has led the economic recovery.



Irish Crisis: Origins and Resolution, Fig. 12 Debt burden (*Source*: CSO National Income and Expenditure)



Irish Crisis: Origins and Resolution, Fig. 13 GNP per head (*Source*: CSO National Income and Expenditure)

As shown in Fig. 13, by 2014 GNP per head had almost returned to the pre-crisis level, and continued growth in 2015 saw that level exceeded. However, substantial long-term damage remains. The level of employment is still significantly below its previous peak, and as a consequence unemployment is well above the pre-crisis level. The major dislocation of the labour market during the crisis was reflected in a return to large-scale emigration.

Given the openness of the economy, Ireland could never have escaped the impact of the great recession that affected the world economy. However, the costs could have been dramatically reduced if wise policies had been pursued over the 2000s. In particular, as argued in Bergin et al. (2011), a gradual tightening of fiscal policy, together with targeted fiscal measures affecting the property market (such as a tax on mortgage interest payments), could have prevented the bubble in the domestic housing market and protected the competitiveness of the economy. The UK Treasury, when it considered the implications for the UK of possible membership of EMU, recommended such a use of targeted fiscal policy action to manage the housing market within EMU (HM Treasury 2003).

Alternatively, or in addition, if appropriate action had been taken by the Central Bank and the financial regulator (a subsidiary of the Central Bank) to prevent the massive expansion of credit to fund investment in property, this could also have prevented the property market bubble. Even if funding for the bubble had been found from sources other than domestic banks (as, for example, in Estonia), appropriate regulatory action would have protected Ireland from the costs ensuing from the collapse of the domestic financial system.

The crisis has led to the largest political upheaval in Ireland in the last 50 years. The main government party (Fianna Fáil) lost the vast bulk of its seats in the 2011 election, something that was unprecedented. Again in the 2016 election the then governing parties (Fine Gael and Labour) lost very heavily. This reflected the fact that the population as a whole were exceptionally angry about the extent of the financial crisis and it also reflected a popular understanding that the origin of the crisis lay in domestic policy mistakes.

While the recovery has reflected the tough fiscal action undertaken by successive governments and the underlying strength of the tradable sector of the economy, it would not have been possible without the support of Ireland's EU neighbours and the IMF. The recovery shows that what Ireland experienced was a liquidity crisis, not a problem of insolvency. The recovery has also been facilitated by the low interest rate environment, which has made the legacy of debt much lighter than was the case with the previous crisis in the 1980s.

Although governments made the mistakes that resulted in huge costs to the economy in the crisis years, the management of that crisis by subsequent successive governments has been quite effective. There can be no doubt now that, while very costly in the short term, there was no alternative to the fiscal adjustment undertaken. It did not seriously damage the underlying fabric of the productive sector of the economy, as evidenced by the return to sustained quite rapid growth since 2012.

While it might be argued that it would have been better to have implemented the adjustment more rapidly, as in Estonia and Latvia, it took some time for Irish policymakers to understand the full severity of the crisis, especially the size of the problem in the banking system. On the other hand, to have delayed the adjustment in Ireland so that the necessary austerity spanned three parliaments, instead of two, would not have been politically sustainable. The Irish electorate has shown its displeasure with the three main political parties in two elections. If further fiscal adjustment had been necessary today it would have been very difficult to implement politically.

The root and branch reform of financial regulation has now been strengthened by the move to the Single Supervisory Mechanism. New macroprudential tools have been developed to manage the exposure of the financial system (and of individuals) to the housing market, and they were deployed in 2015 to prevent the strong recovery getting out of hand. The importance of such

regulation, and its absence in the past, has been highlighted by the Irish crisis. The costs from lax regulation are likely to be dramatically higher than those from over-zealous regulation.

The task of undertaking a large fiscal adjustment in Ireland would have been greatly eased if, at the level of the euro area, a counter-cyclical fiscal policy had been pursued by countries that did not face major financial problems. Instead, fiscal policy in the 2010–2013 period was strongly pro-cyclical (Euroframe 2013). Current EU fiscal rules hold out no prospect of a more appropriate policy response if such a crisis should affect some euro area members in the future (Lane 2015b).

The experience of countries like Estonia and Latvia, which were outside the euro area in 2008, shows that the availability of cheap money allowed an inappropriate expansion of domestic credit independent of EMU membership. EMU was not the cause of the problem; rather unwise domestic policies lead to the crisis in Ireland, Portugal, Spain and Greece. However, the experience of Ireland, in contrast to that of Estonia and Latvia, shows that if a crisis hits, countries receive more generous support when they are members of the euro area.

See Also

- ▶ Credit Crunch Chronology: April 2007–September 2009
- ▶ Euro Zone Crisis 2010
- ▶ Ireland, Economics in

Bibliography

Barrett, A., and A. Bergin. 2009. Estimating the impact of immigration in Ireland. *Nordic Journal of Political Economy* 35: 1–15.

Barrett, A., J. FitzGerald, and B. Nolan. 2002. Earnings inequality, returns to education and immigration into Ireland. *Labour Economics* 9(5): 665–680.

Barrett, A., I. Kearney, and Y. McCarthy. 2007. *Quarterly economic commentary*. Dublin: Economic and Social Research Institute.

Barry, F. 2002. The Celtic Tiger era: Delayed convergence or regional boom? *Quarterly Economic Commentary*. Dublin: Economic and Social Research Institute.

Bergin, A., J. FitzGerald, I. Kearney, and C. O'Sullivan. 2011. The Irish fiscal crisis. *National Institute Economic Review* 217: R47–R59.

Byrne, D., D. Duffy, and J. FitzGerald. 2014. *Household formation and tenure choice: Did the great Irish housing bust alter consumer behaviour?* ESRI working paper 487.

Cardiff, K. 2016. *RECAP: Inside Ireland's financial crisis*. Dublin: Liffey Press.

Conefrey, T., and J. FitzGerald. 2010. Managing housing bubbles in regional economies under EMU: Ireland and Spain. *National Institute Economic Review* 211(1): 211–299.

Crafts, N. 2014. Ireland's medium-term growth prospects: A phoenix rising? *Economic and Social Review* 45(1): 87–112.

Euroframe. 2013. *Economic assessment of the euro area*. Winter report. www.euroframe.org

FitzGerald, J. 2013a. The impact of fiscal policy on the economy. *Quarterly Economic Commentary*. Research notes 2013/3/1.

FitzGerald, J. 2013b. Financial crisis, economic adjustment and a return to growth in the EU. *Revue de l'OFCE – Debates and Policies* 127: 277–302.

FitzGerald, J. 2015. Problems interpreting the national accounts in a globalised economy – Ireland. Special Article. *Quarterly Economic Commentary*. Dublin: Economic and Social Research Institute.

FitzGerald, J., and E. Morgenroth. 2006. *Ex ante evaluation of the investment priorities for the national development plan 2007–2013*. Policy research series no. 59. Dublin: Economic and Social Research Institute.

FitzGerald, J., A. Bergin, Í. Kearney, A. Barrett, D. Duffy, S. Garrett, and Y. McCarthy. 2005. *Medium-term review: 2005–2012*. Dublin: Economic and Social Research Institute.

HM Treasury. 2003. *Fiscal stabilisation and EMU*. London: HM Treasury.

Honohan, P. 2009. Resolving Ireland's banking crisis. *Economic and Social Review* 40(2): 207–231.

Honohan, P. 2010. *The Irish banking crisis: Regulatory and financial stability policy 2003–2008*. Report to the Minister for Finance by the Governor of the Central Bank.

Kelly, M. 2007. On the likely extent of falls in Irish house prices. *Quarterly Economic Commentary*. Dublin: Economic and Social Research Institute.

Lane, P.R. 2015a. The funding of the Irish domestic banking system during the boom. *Journal of the Statistical and Social Inquiry Statistical Society of Ireland* 44: 40–70.

Lane, P. R. 2015b. *Macro-financial stability under EMU*. Trinity economic paper no. 06/15.

O'Gráda, C. 2002. Is the Celtic Tiger a paper tiger? *Quarterly Economic Commentary*. Dublin: Economic and Social Research Institute.

Oireachtas. 2016. *Report of the banking inquiry*. https://inquiries.oireachtas.ie/banking/

Whelan, K. 2014. Ireland's economic crisis: The good, the bad and the ugly. *Journal of Macroeconomics* 39(B): 424–440.

Iron Law of Wages

Mark Blaug

Keywords

Iron law of wages; Lassalle, F.; Market price of labour; Natural price of labour; Natural wages; Population growth; Real wages; Ricardo, D.; Subsistence' theory of wages

JEL Classifications

J3

The 'iron (or brazen) law of wages' is a term invented by Ferdinand Lassalle (1862) to describe the inexorable tendency of real wages under capitalism to adhere to a level just sufficient to afford the bare necessities of life. This law, he claimed, was not just a socialist indictment of capitalism but was authorized by leading 'bourgeois' economists such as Malthus and Ricardo. He failed to point out, however, that in Malthus and Ricardo the so-called 'subsistence' theory of wages was predicated on a theory of population growth according to which the supply of labour responds automatically to any gap between the going 'market price' and 'natural price' of labour, the latter being defined as a real wage sufficient to reproduce a working population of given size and composition. Lassalle, however, being a socialist, followed Marx in rejecting the Malthusian theory of population; what ensured the 'iron law of wages' for Lassalle, as for Marx, was the tendency for any rise in real wages to generate unemployment, thus setting in motion forces that reversed the rise. This threw the entire weight of argument for equilibrium adjustments in the labour market on the side of employers' demand; it provided no explanation of the supply of labour and thus failed to furnish a determinate theory of wages in long-run equilibrium. Ironically, therefore, there may be an 'iron law of wages' in Malthus and Ricardo, but there is certainly no such iron law in socialist economics. The question whether Malthus and particularly Ricardo can be said to have held the iron law or subsistence theory of wages was a favourite debating question in the latter half of the 19th century (see, for example, Marshall 1890, pp. 508–9). There is no doubt that they held the view that real wages tend to fluctuate around a natural point of 'gravity', namely, the minimum level of food and other necessities required for existence. But, in the first place, these fluctuations, depending as they did upon decisions to marry and to have children, involved a lag of at least 15–18 years, a point which Malthus (but not Ricardo) conceded explicitly. In the second place, the minimum-of-existence level of 'natural wages' was admitted to be a matter of custom and habit and therefore subject to a secular upward drift. It was therefore perfectly possible to argue for the existence of something like a normal long-run supply price of labour – a constant real wage, everything else being the same – while at the same time granting that the 'market price' of labour fluctuated around an ever-rising trend. In short, rising living standards under capitalism do not violate the iron law of wages, understood as a theory about the long-run equilibrium price of labour. But that is only to say that the iron law or subsistence theory of wages amounts for all practical purposes to accepting customary wages as an institutional datum (Schumpeter 1954, p. 665).

There has been a revisiting of the old debate about whether Ricardo held the iron law of wages, but in an entirely new form: did Ricardo hold real wages to be constant at the subsistence level in stationary equilibrium or did he allow for an initial stage of increasing real wages followed by a final stage of declining wages alongside a secular fall in the rate of profit (Hollander 1983)? It is doubtful whether this question yields one simple, neat answer, since it is clear that Ricardo operated with a number of different models regarding the determination of the 'natural price' of labour. In the very opening paragraph of the chapter on wages in Ricardo's *Principles of Political Economy of Taxation*, the 'natural price' of labour is defined as 'that price which is necessary to enable the labourers, one with another, to subsist and to perpetuate their race, without either increase or diminution'. This defines the natural price of

labour to be the commodity wage that ensures a zero rate of population growth. But a page or two later, the natural price of labour is said to be that commodity wage which ensures a rate of growth of population equal to the rate of growth of the capital stock, so that market wages only rise above natural wages when capital accumulates faster than the growth of population. It is possible to make sense of this in terms of modern growth theory, and many have done so (see Casarosa 1978), but it is questionable whether Ricardo himself was aware of what he was doing, the more so as he frequently resorts to the constant-subsistence-wage assumption in the later tax chapters of the *Principles*.

See Also

► Wages Fund

Bibliography

Casarosa, C. 1978. A new formulation of the Ricardian system. *Oxford Economic Papers* 30 (1): 38–63.

Hollander, S. 1983. On the interpretation of Ricardian economics: The assumption regarding wages. *American Economic Review* 73: 314–318.

Lassalle, F.J.G. 1862. *Open letter to the National Labor Association of Germany*. Trans. J. Ehmann and F. Bader. Cincinnati: Cincinnati Press, 1879.

Marshall, A. 1890. *Principles of economics*. 9th (Variorum) edn, ed. C.W. Guillebaud, vol. 1. London: Macmillan, 1961.

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

Irreversible Investment

Janice C. Eberly

Abstract

The cost of an irreversible investment cannot be recovered once it is installed. This restriction not only truncates negative investments, but also raises the threshold for positive investment. The threshold return that justifies an irreversible investment increases with uncertainty, or more precisely, with the probability mass in the lower tail of outcomes. Irreversibility constrains the ability to redeploy capital in 'bad' states, so the agent is particularly sensitive to these states when investing *ex ante*.

This finding is analogous to valuation and exercise of financial options, and irreversible investments are valued and understood by using option pricing techniques.

Keywords

Adjustment costs; Irreversible investment; Option pricing theory; Option valuation; Put–call parity; Uncertainty

JEL Classifications

D4; D10

Irreversible investment acknowledges that the value of capital may not be fully recoverable when resold.

This simple generalization has rich implications for investment. Beyond truncating disinvestment, irreversibility changes the dynamics of investment by creating a threshold level of returns for positive investments. Below this threshold, investment is zero – which immediately implies intermittent rather than continuous investment activity. Moreover, the threshold return that justifies investment exceeds the required return on a reversible investment.

Investment and Options

Marschak (1949) raised the potential role of irreversibility in factor accumulation by emphasizing the convertibility or liquidity of capital. Work by Arrow (1968) and Henry (1974) considered when irreversible actions in environmental applications were justified and emphasized the idea of an option value. This idea was extended by Bernanke (1983) to the role of uncertainty in delaying investment decisions.

McDonald and Siegel's (1986) article 'The Value of Waiting to Invest' provides the first explicit valuation of investment allowing for irreversibility, incorporating option valuation (real options) into investment theory. McDonald and Siegel analyse a project of fixed size, so the timing of the project is the only choice to be made. They show that the value of the project includes an 'option value of waiting', that can be valued and interpreted using option pricing theory. The additional value of being able to choose when to invest, rather than a 'now or never' investment decision, can be quantitatively large, and has interesting implications for the investment decision. First, the presence of this option implies that it is optimal to delay the investment, rather than undertaking it immediately, even when immediate execution has positive value. Instead, value can be increased by waiting for additional information. Second, like most options, the value of the option to wait is increasing in uncertainty. This feature implies an effect of uncertainty on the value and timing of investments that is absent in most conventional models.

Later work by Pindyck (1988) and Bertola (1988), allows for incremental investment, so that the firm chooses both the timing and the size of its investments. They show that there is a threshold for investing with irreversibility that exceeds the return that would justify a positive reversible investment. Instead of a single investment decision, as in McDonald and Siegel, there is an infinite sequence of investment decisions, where each satisfies the threshold condition.

An Illustrative Model

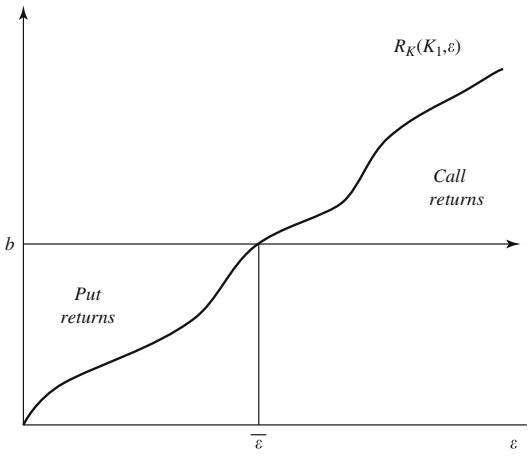
Most irreversible investment models work in continuous time, so that optimal investment timing can be calculated exactly. An introduction to these techniques, as well as a broader overview, is found in Dixit and Pindyck's (1994) *Irreversible Investment*. The intuition can be understood in a discrete time framework, adapted from Abel et al. (1996), specialized to the case of irreversible investment.

Consider the decision of a single firm to undertake a capital investment at time 1. In the first period, the return to installing capital K_1 is $r(K_1)$. The total return $r(K_1)$ is strictly increasing and concave in K and satisfies the Inada conditions. The firm pays a price b per unit of capital to purchase capital. In the second period, the return to capital is uncertain and equal to $R(K, \varepsilon)$, where ε is stochastic. The derivative of $R(K, \varepsilon)$ with respect to K,R_K $(K, \varepsilon) \geq 0$, is continuous and strictly decreasing in K, continuous and strictly increasing in ε, and $R(K, \varepsilon)$ also satisfies the Inada conditions. Define a threshold value of ε by

$$R_K(K_1, \overline{\varepsilon}) = b, \tag{1}$$

as illustrated in Fig. 1.

Assume that the resale price of capital is zero, or complete irreversibility. In the second period, the capital stock is optimally chosen at a level equal to $K_2(\varepsilon)$, subject to the irreversibility constraint. When $\varepsilon > \overline{\varepsilon}$, the optimal capital stock rises to satisfy the first-order condition $R_K[K_2(\varepsilon), \varepsilon] = b$. However, when $\varepsilon > \overline{\varepsilon}$, the marginal return to capital is less than its purchase price. If the firm could resell capital at its acquisition price b (costless reversibility) it would do so. However, the available resale price is zero, so the firm prefers to keep its capital stock, which has positive marginal return; in this case $K_2(\varepsilon) = K_1$. The optimal second-period marginal return to capital is graphed in Fig. 1 as the lower envelope of $R_K(K_1, \varepsilon)$ and b.



Irreversible Investment, Fig. 1 The second period marginal return to capital

Conditional on the optimal second-period capital stock, the firm chooses its capital stock at time 1 to maximize $V(K_1) - bK_1$, where $V(K_1)$ is the first period value of the firm equal to $r(K_1)) \gamma E[R(K_2, \varepsilon)]$ and $0 < \gamma < 1$ is the discount factor. The first-order condition for the optimal capital choice is

$$V'(K_1) \equiv r'(K_1) + \gamma \int_{-\infty}^{\bar{\varepsilon}} R_K(K_1, \varepsilon) \mathrm{d}F(\varepsilon) + \gamma b[1 - F(\bar{\varepsilon})] = b, \tag{2}$$

where $F(\varepsilon)$ is the cumulative distributive function (CDF) of ε.

Notice that the term $V'(K_1)$ is the marginal value of an additional unit of capital, or marginal q. The standard investment first-order condition equating marginal q to the marginal cost of capital still holds with irreversibility. The effects of irreversibility are incorporated into the value of marginal q, so when investment is non-zero the standard q-theory first-order condition equating the marginal value and the marginal cost of investment still holds.

Embedded Options

Now rewrite this first-order condition to highlight the investment options and their implications for the investment decision. Rewrite Eq. (2) as

$$q(K_1) \equiv V'(K_1) \equiv n(K_1) - \gamma c(K_1) \tag{3}$$

where

$$n(K_1) \equiv r'(K_1) + \gamma \int_{-\infty}^{\infty} R_K(K_1, \varepsilon) \mathrm{d}F(\varepsilon) > 0 \tag{4}$$

and

$$c(K_1) \equiv \int_{\bar{\varepsilon}}^{\infty} [R_K(K_1, \varepsilon) - b] \, \mathrm{d}F(\varepsilon) > 0. \tag{5}$$

The marginal value of an additional unit of capital is decomposed into two terms. The first term, $n(K_1)$, is equal to the present value of marginal returns to capital, evaluated at its current level, K_1. The second term subtracts the discounted value of a call option, $c(K_1)$, to add more capital, as illustrated in Fig. 1, where the returns to the call option are represented by the area under $R_K(K_1, \varepsilon)$ and above the line b. The call option reduces the marginal value of capital because additional capital irrevocably reduces the marginal return to capital owing to the concavity of the revenue function. If one combines these two terms, the marginal value of capital is the discounted sum of marginal revenues on the assumption that the capital stock is fixed, less the marginal value of the option to increase the capital stock. Note that the concavity of the revenue function is crucial to this mechanism. Hence, models such as Abel and Eberly, (1997) which assume constant returns to scale, do not generate these option values.

The effects of uncertainty are not transparent in the above formulation, since both terms in Eq. (3) depend on the distribution, $F(\varepsilon)$. To better discern the effect of uncertainty, rewrite Eq. (2) instead as

$$q(K_1) \equiv V'(K_1) \equiv j(K_1) - \gamma p(K_1) \tag{6}$$

where

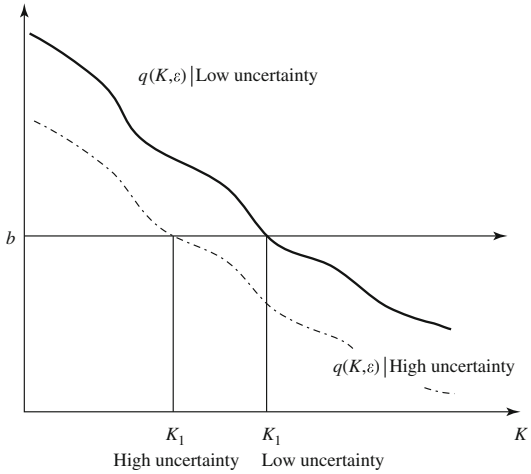
$$j(K_1) \equiv r'(K_1) + \gamma b > 0 \tag{7}$$

and

$$p(K_1) \equiv \int_{-\infty}^{\bar{\varepsilon}} [b - R_K(K_1, \varepsilon)] \, \mathrm{d}F(\varepsilon) > 0. \tag{8}$$

The marginal value of an additional unit of capital is again decomposed into two terms. The first term, $j(K_1)$, is the discounted marginal return to costlessly reversible capital: the firm earns the marginal revenue in period one and can sell the capital for the same price b in period two. This is the Jorgensonian marginal return (Jorgenson, 1963); notice that it is independent of ε and risk free. The second component of q is the put option to sell capital at price b. When investment is irreversible, the put option is not available to the firm, since it cannot sell capital at any

positive price. The value of the put option must be subtracted from the Jorgensonian valuation (where resale at price b would be permitted) to obtain the marginal value of irreversible capital. Marginal q can thus be written as a frictionless value less the value of the put option that is eliminated by the irreversibility constraint. This is illustrated in Fig. 1 by subtracting the returns to the put option (the area under the line b and above the function $R_K(K_1, \varepsilon)$) from the frictionless return b.



Irreversible Investment, Fig. 2 Marginal q and the optimal capital stock under low and high uncertainty

Effects of Uncertainty and Put–Call Parity

To calculate the effect of uncertainty on marginal q from Eq. (6), one need only calculate the effect of uncertainty on $p(K_1)$, since $j(K_1)$ is risk free. The effect of uncertainty on $p(K_1)$ is clear: $p(K_1)$ is an option value, and an increase in uncertainty increases the value of an option. In this case specifically, a second-order stochastic dominant shift in the distribution of ε shifts the CDF up for every value of ε. Since $R_K(K_1, \varepsilon)$ is increasing in ε, the term $[b - R_K(K_1, \varepsilon)]$ is decreasing in ε. Hence, greater uncertainty in ε shifts more weight of the CDF towards the large option payoffs in the left tail and unambiguously increases the value of the option, $p(K_1)$. Greater uncertainty unambiguously lowers the value of $q(K_1)$. Since $q(K_1)$ is decreasing in K_1, a downward shift in $q(K_1)$ reduces the optimal value of K_1 for a given value of b, as illustrated in Fig. 2. This decrease is the incremental investment counterpart to McDonald and Siegel's finding that greater uncertainty increases the option value of waiting, lowering the value of investing immediately.

This formulation of $q(K_1)$ also demonstrates Bernanke's (1983) 'bad news principle' of irreversible investment. The distribution of ε only appears in the expression for q in Eq. (6) via the put option $p(K_1)$. The put option only depends on the lower tail of the distribution of ε, below the threshold $\bar{\varepsilon}$. That is, the only part of the distribution of shocks that affects the value of $q(K_1)$ is the lower tail – or the 'bad news'. The upper tail is irrelevant, since in that region, the firm invests until the marginal product of capital equals its price. The exact realization of the shock in this region is irrelevant to the marginal return. In the lower tail, on the other hand, the firm neither invests nor disinvests, and the realization of the shock determines the marginal return to capital.

Figure 1 illustrates these arguments. The second period return to capital is the lower envelope of the price of capital, b, and the second-period marginal return, $R_K(K_1, \varepsilon)$ evaluated at K_1. The value of these returns depends on ε only in the lower tail of the distribution of ε (the bad news principle). The lower envelope can be expressed as either the function $R_K(K_1, \varepsilon)$ less the area labelled *call returns* in Fig. 1; adding this difference to first-period marginal returns $r'(K_1)$, we obtain the expression for q in Eq. (3). Equivalently, the second-period marginal return can be expressed the line b less the area labelled *put returns* in Fig. 1. Adding the first-period marginal return $r'(K_1)$, we obtain the expression for marginal q in Eq. (6). The fact that the second-period return can be written in two equivalent ways using options follows from put–call parity, a fundamental property of options prices. In fact, in this setting put–call parity is found simply by setting the two expressions for q in Eqs. (3) and (6) equal to each other. Equating these two expressions for q and simplifying, we find

$$\gamma p(K_1) + n(K_1) = \gamma c(K_1) + j(K_1). \quad (9)$$

This expression equates the value of a portfolio containing a put option and the underlying security, $n(K_1)$, to the value of a portfolio containing a call option and a risk-free asset. For a financial security such as a stock with price S, put–call parity analogously states that $P(S, \tau) + S = C(S;, \tau) + X/(1+r)^{\tau}$, where X is the strike price of the options and τ is the time to maturity. The terms $P(S, \tau)$ and $C(S, \tau)$ are the value of the put and call, respectively, on the underlying stock, S. $X/(1 + r)^{\tau}$ is the present value of a risk-free payoff (a zero coupon bond) of X in τ periods.

Extensions and Applications

The above analysis assumes complete irreversibility. However, less stringent forms of the constraint deliver similar implications. Abel and Eberly (1996) examine costly reversibility, where capital can be disinvested and resold at a price less than the purchase price of capital. In this case, the gap between the investment and disinvestment thresholds opens quickly, even for small differences between the purchase and sale prices of capital. Moreover, this formulation has assumed kinked, linear adjustment costs, so that the degree of irreversibility is summarized by the ratio of the purchase and sale prices of capital. However, with more general cost formulations, such as Abel and Eberly (1994), capital may have a positive resale price and still be effectively irreversible when other costs of reselling capital exceed any potential benefits. In addition to a resale market discount, convex adjustment costs and fixed costs, for example, may induce irreversibility.

Research on irreversibility has branched out both empirically and theoretically. Initial applications included energy and natural resource markets (Brennan and Schwartz, 1985), with extensions to virtually all types of quasi-fixed capital, including durable goods, real estate and equipment investment. Modelling has been extended to include multiple types of quasi-fixed capital goods (Eberly and van Mieghem, 1997). Aggregating models with infrequent adjustment to incorporate equilibrium effects is challenging, and the results remain controversial. Except in very special cases (Caplin and Spulber, 1987) aggregating requires tracking a distribution of agents. However, it is precisely this feature that can match the observation that much of the volatility in empirical investment arises from the extensive margin (the number of agents adjusting) rather than the intensive margin (the average size of the adjustment). Much progress has been made in this direction (for example, Caballero and Engel, 1999), though the quantitative implications vary with modelling strategy (Veracierto, 2002).

See Also

- ▶ Adjustment costs
- ▶ Marschak, Jacob (1898–1977)
- ▶ s-S models
- ▶ Tobin, James (1918–2002)
- ▶ Tobin's q

Bibliography

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

Abel, A.B.., and J.C. Eberly. 1996. Optimal investment with costly reversibility. *Review of Economic Studies* 63: 581–593.

Abel, A.B.., and J.C. Eberly. 1997. An exact solution for the investment and value of a firm facing uncertainty, adjustment costs, and irreversibility. *Journal of Economic Dynamics and Control* 21: 831–852.

Abel, A.B.., A.K. Dixit, J.C. Eberly, and R.S. Pindyck. 1996. Options, the value of capital, and investment. *Quarterly Journal of Economics* 111: 753–777.

Arrow, K.J. 1968. Optimal capital policy and irreversible investment. In *Value, capital, and growth*, ed. J.N. Wolfe. Chicago: Aldine.

Bernanke, B.S. 1983. Irreversibility, uncertainty, and cyclical investment. *Quarterly Journal of Economics* 98: 85–106.

Bertola, G. 1988. Adjustment costs and dynamic factor demands: Investment and employment under uncertainty. Ph.D. Dissertation, Cambridge, MA: Massachusetts Institute of Technology.

Brennan, M., and E. Schwartz. 1985. Evaluating natural resource investments. *Journal of Business* 58: 135–157.

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

Caplin, A.S., and D.F. Spulber. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102: 703–726.
Dixit, A.K., and R.S. Pindyck. 1994. *Irreversible investment*. Princeton: Princeton University Press.
Eberly, J.C., and J.A. van Mieghem. 1997. Multi-factor dynamic investment under uncertainty. *Journal of Economic Theory* 75: 345–387.
Henry, C. 1974. Option values in the economics of irreplaceable assets. *Review of Economic Studies* 41: 89–104.
Jorgenson, D. 1963. Capital theory and investment behavior. *American Economic Review* 53: 247–259.
Marschak, J. 1949. Role of liquidity under complete and incomplete information. *American Economic Review* 39: 182–195.
McDonald, R.L., and D. Siegel. 1986. The value of waiting to invest. *Quarterly Journal of Economics* 101: 707–728.
Pindyck, R.S. 1988. Irreversible investment, capacity choice, and the value of the firm. *American Economic Review* 78: 969–985.
Veracierto, M.L. 2002. Plant-level irreversible investment and equilibrium business cycles. *American Economic Review* 92: 181–197.

Islamic Economic Institutions

Timur Kuran

Abstract

The economic institutions of the classical Islamic world include Islamic contract law and the waqf, a form of trust. Until modern times, these two institutions were generally beneficial to economic performance. However, each had limitations that eventually blocked modern economic growth. Islamic contract law discouraged the formation of large and long-lived partnerships, thus obviating the need for business techniques and organizational forms associated with economic modernization. The waqf, designed as a rigid organization, locked capital into inefficient uses. Not until modern times has the corporation, a more flexible organizational form, entered the legal systems of the Islamic world.

Keywords

Charitable contributions; Choice of law; Contract law; Corporation; Double-entry bookkeeping; Industrial revolution; Inheritance; Interest; Islamic economic institutions; Limited liability; Partnerships; Tax farming; Waqf; Zakat

JEL Classifications

N4

Prior to the eighteenth century, the Islamic world did not appear economically underdeveloped to outside observers. Comparative studies by economic historians confirm that it became 'poor' in relation to Europe during the Industrial Revolution. Until that point, economic institutions grounded in Islamic law had afforded a respectable level of wealth by standards of the day. They had also facilitated the spread of Islam across Asia, southern Europe, and the coasts of Africa.

Law of Contracts

The first few centuries of Islam – *c.* 622–1000 AD – witnessed the gradual development of an elaborate law of contracts. It enabled the pooling of labour and capital through several forms of partnership, including ones providing limited liability to passive investors. Profit shares, negotiated in advance, could be unequal or contingent. Islamic partnership contracts were enforced, with minor variations, wherever Muslims ruled. As merchants and producers moved, they carried Islamic law with them, helping to spread Islam. Huge numbers of people converted in order to gain acceptance into lucrative commercial networks managed according to Islamic law.

Islamic law limited neither the size of a partnership nor its duration. However, in practice the typical Islamic partnership consisted of two people, who pooled resources for a single economic venture expected to last just a few months (Çizakça 1996). Lacking a life of its own, it was not what we call a firm. If a partner died during the contract period, the partnership became null and

void, and the decedent's share of the assets fell to his heirs. There could be numerous claimants, for the Islamic inheritance system, by medieval standards remarkably egalitarian, assigns mandatory shares to a possibly long list of extended relatives. Accordingly, reconstituting a dissolved partnership could be very costly. Merchants and investors minimized the risk of dissolution by keeping their partnerships small and ephemeral (Kuran 2003b).

A long-term consequence is that Islamic partnerships remained structurally simple, which obviated pressures to develop the sorts of organizational forms and business techniques that, in western Europe, gradually led to the modern economy. For instance, double-entry bookkeeping did not develop, and no markets arose for trading enterprise shares. This institutional inertia made it impossible to borrow new organizational forms, except as part of a comprehensive legal reform. Advanced organizational forms, such as the joint-stock company and the corporation, reached the Islamic world in the nineteenth century through the imposition of secular commercial law. By that time the financing and organization of the region's external trade was largely under Western control; and, as a result of the Industrial Revolution, productivity was much higher in the West than elsewhere. The very commercial institutions that had served Muslims well through the Middle Ages were now hindering the exploitation of modern technologies.

Role of Minorities

The religious minorities of the Islamic world might have escaped the limitations of Islamic commercial institutions, because they enjoyed 'choice of law' – the privilege to do business under legal systems of their own. Yet as individuals non-Muslims could opt unilaterally to take anyone to an Islamic court, whose decision would trump that of a non-Muslim judge or arbitrator. To achieve predictability in their economic relations, non-Muslims thus tended to base their financial and commercial contracts on Islamic law; their claims induced their own court systems to emulate Islamic legal practices. Consequently, until the eighteenth century the economic performance of non-Muslim peoples of the Islamic world did not diverge significantly from that of Muslims. Most non-Muslim communities started pulling ahead, however, as western Europe developed the legal infrastructure of modern capitalism. Vast numbers of Christians, Jews and other non-Muslims gained an economic advantage over Muslims by doing business under western or western-inspired laws (Kuran 2004b; Issawi 1982).

The Waqf

Another contributor to the Islamic world's economic successes and also to its subsequent economic retardation is the waqf, Islam's distinct form of trust. From the eighth century to modern times, Muslim-governed states provided few public goods directly, beyond law and order. They left the supply of public goods largely to waqfs established in a decentralized manner. Vast resources flowed into waqfs; by the early eighteenth century they owned between a quarter and half of all real estate, depending on the country. The services financed through waqfs included mosques, schools, hospitals, water fountains, roads, parks, inns, bathhouses, orphanages and soup kitchens.

A waqf is an unincorporated trust established under Islamic law by an individual owner of immovable property for the perpetual provision of a service. It emerged in the early Islamic period, a time of weak property rights, partly to enable landowning high officials to shelter wealth. Converting property into waqf yielded considerable immunity against confiscation, because waqf-owned assets were considered sacred, and this made legitimacy-seeking rulers reluctant to expropriate them. In addition to social status and religious satisfaction, the founder usually obtained pecuniary benefits. He could make himself the waqf's mutawalli (trustee and manager), set his own salary, appoint relatives to paid positions, and designate his successor. This last prerogative enabled circumvention of the Islamic inheritance system. In founding a waqf, then, an individual did not simply engage in charity. In

return for shouldering social responsibilities, he obtained the privilege of sheltering wealth for personal use. Local norms determined the share of a waqf's income that its mutawalli could reserve for himself and his family.

For a millennium this system for supplying public goods remained a distinguishing feature of the Islamic world. It owed this remarkable longevity to identifiable benefits that it yielded to huge groups. Property owners achieved a measure of material security. Rulers unburdened themselves of the responsibility to provide public goods. And the average person received diverse forms of philanthropy. Nevertheless, the waqf system had a flaw that became increasingly serious over time. Although some opportunities existed to reallocate resources to new uses, the waqf was designed to serve its founder's wishes for ever. As such, it could not adapt quickly to changing social needs, and it locked capital into inefficient uses. By the nineteenth century, a time of massive technological change, the waqf system had become conspicuously dysfunctional, and reformers took to dismantling it (Çizakça 2000; Kuran 2001).

Up to that time, services to the Middle East's great cities were supplied mostly by waqfs. The nineteenth century saw the establishment of the region's first municipalities, under secular laws. These municipalities, which attained corporate powers, could reallocate resources relatively quickly. Within a few decades, they assumed most of the functions previously relegated to the waqf sector.

Absence of the Corporation

Islamic law, which borrowed from various pre-existing legal systems, had spurned the Roman concept of the corporation. Limiting legal standing to natural persons supported Islam's political mission, which was to turn Arabia's feuding tribes into an undivided religious community. Corporations might have undermined that goal by enabling tribes to form autonomous organizations. During the formative period of Islam – from the seventh through the tenth centuries – the Middle East thus experienced no incorporation wave analogous to that observed in contemporaneous western Europe. One reason is that the waqf, by providing the means for delivering perpetual services with large sunk costs, alleviated the need for corporations. Another is that the waqf system spawned constituencies with a stake in preserving its key features; yet another that merchants and producers who stood to benefit from corporate powers could not muster the collective action necessary to reform the legal system. Not until the modern era did the concept of the corporation enter legal systems of the Islamic world.

In the late twentieth century, certain predominantly Muslim countries started to revive their waqf sectors, though in modernized and secularized form. Unlike the traditional waqf, a modern waqf enjoys legal personhood, and its founder may be a group. It is managed by a mutawalli board rather than a single caretaker appointed for life. Most critical, as a self-governing organization it can remake itself. Secularists, civil rights groups and economic liberalizers are the key constituencies of coalitions formed to promote waqf founding. These groups see their mission as a vehicle for shrinking the state, strengthening local governance, and promoting democratization (Çizakça 2000). Thus, having played an enormous role in Islamic economic history, the waqf is now turning into an agent of political and economic modernization.

Because the Qur'an does not mention the waqf, many Islamists are indifferent to ongoing efforts to reinvigorate the waqf sector. Their overriding goal is to purge interest from financial transactions, largely in the belief that the Qur'an bans interest categorically (Saleh 1986; Lewis and Algaoud 2001; Kuran 2004a). In fact, Islam's prescriptions concerning interest have always been a matter of interpretation, and throughout Islamic history interest-based transactions have been common (Rodinson 1966). Nevertheless, Islamists treat Islamic banking, intended to be free of interest, as the sine qua non of a properly Islamic economy. Yet Islamic banking is a modern creation. Pre-modern economies based on Islamic law had moneylenders but no banks (Udovitch 1979).

The first banks of the Islamic world, all foreign-owned and -managed corporations, date from the mid-nineteenth century (Kuran 2005).

Property Rights

Until modern times economies of the Islamic world suffered from a lack of institutions to tie the hands of governments. This meant that private property rights remained weak. Although material insecurity varied across time and space, taxation was often arbitrary, and states resorted to compulsory labour. Private property rights did not achieve credibility even in the eyes of state officials – one reason why endowing waqfs was so popular. A scribe could be plucked out of obscurity to become a prosperous statesman, and then, all of a sudden, fall into disgrace and lose everything. The expropriation of large estates was especially common, all the more so in times of financial crisis. Because this practice violated the Islamic law of inheritance, typically it was based on the ground that the deceased was not the rightful owner of his estate (Findley 1989).

In the seventh century, the first Islamic state in Arabia had instituted a tax-and- subsidy system that might have strengthened property rights. Known as zakat, it required the payment of taxes to the state in specific forms of income and wealth at predetermined rates. In providing the state the resources to fund various activities, including charity, it also capped taxation (Rahman 1974). However, precisely because of the inflexibility of its rate structure, within a couple of generations revenue-hungry rulers abandoned it for taxes that gave them greater latitude. Thereafter zakat metamorphosed into a narrow religious duty, incumbent on people of means, to assist the poor on an annual basis (Kuran 2003a). Modern Islamists have tried to turn zakat into a state-run social welfare system to which the wealthy make obligatory contributions. But throughout the Islamic world taxation remains an essentially secular matter. It has also become more predictable. The wealthy classes of the present have far better defences than those of the past against government predation.

Taxation can be arbitrary without being devoid of logic. In pursuing opportunities to raise revenue, rulers sought to limit transaction costs, in particular to minimize the costs of measuring income, identifying assets, and collecting taxes. To that end they tended to collect fixed taxes directly, leaving the collection of variable taxes to local officials (Coşgel and Miceli 2005). They also made extensive use of tax farming, which assigns collection rights to people knowledgeable about tax units.

See Also

- ▶ Corporations
- ▶ Development Economics
- ▶ Institutional Trap
- ▶ Public Goods
- ▶ Religion and Economic Development

Bibliography

Çizakça, M. 1996. *A comparative evolution of business partnerships: The Islamic world and Europe, with special reference to the Ottoman archives*. Leiden: E.J. Brill.

Çizakça, M. 2000. *A history of philanthropic foundations: The Islamic world from the seventh century to the present*. Istanbul: Boğaziçi University Press.

Coşgel, M., and T. Miceli. 2005. Risk, transaction costs, and tax assignment: Government finance in the Ottoman Empire. *Journal of Economic History* 65: 806–821.

Findley, C. 1989. *Ottoman civil officialdom: A social history*. Princeton: Princeton University Press.

İnalcık, H. 1994. The Ottoman state: Economy and society, 1300–1600. In *An economic and social history of the Ottoman Empire, 1300–1914*, ed. H. İnalcık and D. Quataert. Cambridge: Cambridge University Press.

Issawi, C. 1982. The transformation of the economic position of the *Millets* in the nineteenth century. In *Christians and Jews in the Ottoman Empire*, ed. B. Braude and B. Lewis, vol. 1. New York: Holmes and Meier.

Kuran, T. 2001. The provision of public goods under Islamic law: Origins, impact, and limitations of the waqf system. *Law and Society Review* 35: 841–897.

Kuran, T. 2003a. Islamic redistribution through *zakat*: Historical record and modern realities. In *Poverty and charity in Middle Eastern contexts*, ed. M. Bonner, M. Ener, and A. Singer. Albany: State University of New York Press.

Kuran, T. 2003b. The Islamic commercial crisis: Institutional roots of economic underdevelopment in the Middle East. *Journal of Economic History* 63: 414–446.
Kuran, T. 2004a. *Islam and Mammon: The economic predicaments of Islamism*. Princeton: Princeton University Press.
Kuran, T. 2004b. The economic ascent of the Middle East's religious minorities: The role of Islamic legal pluralism. *Journal of Legal Studies* 33: 475–515.
Kuran, T. 2005. The logic of financial Westernization in the Middle East. *Journal of Economic Behavior and Organization* 56: 593–615.
Lewis, M., and L. Algaoud. 2001. *Islamic banking*. Cheltenham: Edward Elgar.
Rahman, F. 1974. Islam and the problem of economic justice. *Pakistan Economist* 14(24 August): 14–39.
Rodinson, M. 1966/1973. *Islam and Capitalism.* Trans. B. Pearce. New York: Pantheon.
Saleh, N. 1986. *Unlawful gain and legitimate profit in Islamic law: Riba, gharar, and Islamic banking*. Cambridge: Cambridge University Press.
Udovitch, A. 1979. Bankers without banks: Commerce, banking, and society in the Islamic world of the Middle Ages. In *The Dawn of modern banking*, ed. Center for Medieval and Renaissance Studies. New Haven: Yale University Press.

Islamic Finance

Mahmoud El-Gamal

Abstract

Islamic laws on financial matters date back to medieval times. They were reinterpreted in the 20th century to provide guidelines for the burgeoning Islamic financial sector. Compliance with religious law is a driving force in this sector, and a variety of financial instruments have been developed that are adjudged to be acceptable for use by Muslims. These continue to evolve, and the rules could be considered to merit reinterpretation to better enable Islamic financial institutions to deal with risk factors and obey the spirit, rather than merely the letter, of medieval Islamic jurisprudence, which was regulatory in nature.

Keywords

Islamic finance; *Shari'a* law; *Mudaraba*

JEL Classifications

Z12; G00

The notion of 'Islamic finance' was born during the tumultuous identity-politics years of the mid-20th century. Indian, Pakistani and Arab thinkers contemplated independence from Britain, and the independence of Pakistan from India, within a context of 'Islamic society'. Islam was assumed to inspire political, economic and financial systems that are distinctive and independent of the Western (capitalist) and Eastern (socialist) models of the epoch. The term 'Islamic economics' was coined by Abu al-A'la Al-Mawdudi, whose students and followers worked to develop an ostensible Islamic social science (Kuran 2004). Mawdudi's influence on Arab Islamists began with the writings of Sayid Qutb, the father of modern Arab political Islam, whose quasi-exegesis *Under the Qur'anic Shade* referred exclusively to Mawdudi's writings on economic matters. Mawdudi's migration from majority-Hindu Indian society to majority-Muslim Pakistan thus became a prototype for Islamist migration away from secular political and economic systems.

From Islamic Economics to Islamic Banks

In the first few decades of its existence, Islamic economics focused on comparative economic systems (a fashionable field at the time) as well as neoclassical and Keynesian modelling with a highly stylized *homo islamicus* (a moral and ethical individual who shuns excessive greed and consumerism) in place of mainstream economics' *homo economicus* (a selfish utility and profit maximizer) (Haneef 1995).

As a byproduct, Islamic banking emerged in the Islamic economists' literature as a financial system based exclusively on profit-and-loss sharing, which was argued to be more equitable and stable (Chapra 1996; Siddiqi 1983). In the process, Islamic economists focused on the Islamic prohibition of *riba* or usury, which they interpreted as a prohibition of all interest-based lending, in accordance with earlier interpretations of the Judeo-Christian canon.

Classical Islamic jurisprudence had interpreted interest-based lending, the cornerstone of fractional-reserve depositary banking, as riskless – and therefore illegitimate and inequitable – return for idle capitalists. Indeed, the importance of credit and counterparty risk for any financial analysis remains conspicuously absent from the writings of the Islamic-economics faithful. The preferred financial model, they postulated, would be based on the ancient silent-partnership model known in Islamic writings as *mudaraba*, corresponding to the Jewish *heter iska* and the Christian-European *commenda* (Udovitch 1970).

An 'Islamic bank' was envisioned as a two-tier silent partnership. Thus, deposits seeking a return (as opposed to fiduciary deposits, for which 100 per cent reserves are required) would not be guaranteed loans to the bank, but rather silent-partnership investments in the bank's portfolio. In turn, the bank's investments of those funds would not consist of loans and acquisition of debt instruments, but rather profit-and-loss sharing investments in other silent partnerships. Thus, the Islamic bank would serve its financial intermediation function (pooling of return-seeking savings and diversification of investments) through profit-and-loss sharing. This idea continues to serve as the cornerstone of Islamic banking today, despite being thoroughly debunked by prominent jurists (Tantawi 2001; El-Gamal 2003).

Potential loss of return-seeking deposits was assumed by Islamic-banking proponents such as the Islamic Financial Services Board (IFSB) to encourage depositor-monitoring and risk-mitigating market-based discipline. Thus, the grossly inadequate depositor-protection measures supported by the industry have focused on transparency of operations and profit-distribution mechanisms (IFSB 2006).

The Practice of Islamic Banking

This risk-sharing model has continued to shape the liabilities side of Islamic banks' balance sheets, with a few exceptions in Europe and the United States, where regulators have required Islamic financial providers that function as banks to guarantee deposits. The assets side of Islamic banks and financial providers, on the other hand, has utilised multiple structured-financial models to replicate loans and fixed-return securities that limit the banks' exposure to credit risk. The transformation from the idealistic profit-and-loss sharing model of Islamic economics – which continues to be hailed as the 'Islamic ideal' by industry practitioners and commentators – to replication of modern financial products and markets in 'Islamic' garb coincided with the increased importance of classical methods of Islamic jurisprudence and a limited rhetorical role for Islamic economics.

Early models in the subcontinent during the 1950s and in Egypt during the 1960s notwithstanding, the true beginnings of Islamic banking and finance occurred in the mid-1970s. Islamic jurists including the Shiite scholar Baqir al-Sadr and many Sunni scholars in Egypt, Saudi Arabia and elsewhere collaborated with Islamist bankers to replicate loans using ancient contract forms. Baqir al-Sadr, in his classical work *The Non-Usurious Bank in Islam* (long out of print), had attempted to use similar structured products to replicate guaranteed bank deposits on the liabilities side. However, since risk-sharing depositors were clearly beneficial to the shareholders of Islamic banks, and because the latter drove innovation in Islamic banking through the retention of lawyers and religious scholars, most of the 'innovations' were restricted to the assets side of the balance sheet.

Murabaha (Cost-Plus Sale) Financing

The workhorse of Islamic banking has been the *murabaha* (cost-plus sale) contract. The logical evolution of this form of finance is indicative of the general methodology of Islamic finance to this day. In the early 1980s, Islamic banks in the Gulf were flush with petrodollars, and Western corporations were eager to borrow from them as Western-bank credit dried up following the petrodollar-driven Latin American debt crisis. Islamic banks resorted to the easiest ancient trick: introducing a property to separate lent principal from repaid principal plus interest.

In the simplest ruse, the bank could have sold some commodity to its potential borrower on credit (for principal plus interest payable later), and then bought it back for cash (principal paid immediately), thus effectively replicating the cashflows of the loan, with the commodity making a round trip from bank to customer and back. However, this ancient ruse was forbidden by name as same-item sale resale (*bay' al-'ina*). In practice, one credit sale and one spot sale of liquid commodities were still used to accomplish the desired goal by conducting the second spot sale with a third party.

Interestingly, Al-Rajhi Investment Company in Saudi Arabia, which has one of the strictest religious-scholar boards, received a question on the legitimacy of the credit sale of gold, and ruled that such sales were disallowed because gold is a monetary commodity. Promptly thereafter, the same board was asked if platinum can be sold on credit, and issued a *fatwa* that this was permitted. Thus, Islamic banks could simply trade precious metals, acquiring an amount of platinum (or other metal excluding gold and silver) equal in value to the desired loan principal. The metal was then sold on credit to the Western borrower under a *murabaha* contract, with a credit price equal to the desired principal plus interest. The customer was then able to sell the metal quickly to receive the desired borrowed principal, perhaps less a small transaction cost.

This was the juristic solution first popularised by the late banker Sami Humud in his book *Evolving Banking Transactions in Accordance with Islamic Law* (1976). The prohibition of *riba* (usury) in the Islamic canon and subsequent juristic analysis left room for such ruses. The Qur'an merely mentioned *riba* in the abstract without specifying precisely which transactions were thus forbidden. The Prophetic tradition merely listed six commodities: gold, silver, dates, wheat, barley and salt, all of which were used at some point as commodity monies in the ancient world, stipulating that those may be traded only hand-to-hand and in equal amounts measured by weight or volume. One school of jurisprudence (*Hanafi*) expanded the prohibition to all commodities measured by weight or volume, but still did not treat them as money. Therefore, while trading platinum now for platinum later, or trading gold now for silver later, would both be deemed impermissible based on the *Hanafi* interpretation, trading platinum now for dollars later was considered permissible.

Interestingly, the *Halacha*, developed by Jewish scholars prior and in parallel to the development of Islamic *Fiqh*, forbade such embedded-interest credit sales (Reisman 1995, p. 112). In contrast, all major schools of Islamic jurisprudence (four Sunni and four Shiite) have allowed credit sales at prices possibly exceeding the spot price. Initially, this was only a method for seller financing. Thus, the financier needed first to acquire the property before selling it on credit. In addition, to give the contract an Islamic flavour, the industry adopted the name of an ancient cost-plus sale – *murabaha*, a contract devised to protect buyers who were unfamiliar with market prices, allowing them to negotiate prices by negotiating markup over revealed cost.

The contract that emerged in the 1970s was formally known as 'cost-plus sale to the customer who ordered the initial purchase' (*murabaha lil-'amir bil-shira*'). It was initially subject to scholarly controversy, especially as bankers added provisions to eliminate all forms of risk other than customer credit. In order to eliminate property-related risks, which were ironically the basis on which jurists allowed earning a return on the transaction, they allowed banks to stipulate that the eventual buyer must guarantee to buy the property on credit once the bank acquires it. Eventually, wide consensus emerged and the contract became the workhorse of Islamic banking practices, from large multi-million-dollar loans to Western corporations to retail-bank secured lending.

Tawarruq (Monetization) Financing

In order to reduce transaction costs, especially for retail customers who wished to borrow cash, Islamic banks in Gulf Cooperation Council (GCC) countries revived another ancient financial trick: monetization. This transaction is very

similar to the cost-plus commodity-sale finance model, with the added complication that the Islamic bank executes all three legs of the transaction: (i) buying the principal's worth of metals at the spot price, (ii) selling said metals to the customer on credit for principal plus interest, and (iii) selling the metals back to the dealer, as the customer's agent, for the spot price less a small fee. All three transactions can be concluded within minutes via fax.

This transaction avoids the forbidden two-party sale–resale trick by adding not only one commodity as a degree of separation between lent principal and repaid principal plus interest, but also a third-trading-party degree of separation (the metals dealer) so that every two parties formally trade the commodity only once. The commodity still completes a round-trip (dealer → bank → customer → dealer), spot cash in the amount of desired principal completes one trip (bank → dealer → customer), and the credit-sale-price payment of principal plus interest occurs in the future (customer → bank). This three-party variation on same-item sale resale was also known in ancient and medieval practice, and deemed forbidden or reprehensible by most schools of law. Some medieval scholars within the *Hanbali* school of jurisprudence, which is dominant in the GCC, had permitted this practice. Despite the fact that the most respected 14th-century scholars ibn Qayim and ibn Taymiya forbade the transaction (as merely an expensive and potentially more hazardous type of usury/*riba*), contemporary *Hanbali* jurists who dominate one juristic council in Saudi Arabia permitted the practice in 1998. The same council later forbade the organised practice of Islamic banks using this contract in 2003, but the practice continued to thrive.

Ijara (Lease) Financing, Securitization and *sukuk* (Islamic Bonds)

Despite juristic approval of credit-sale-based financing methods, the practice remained suspect in scholarly as well as general Islamist circles. In addition to objections that the practice merely replicated interest-based financing with interest characterised as profit or markup, there were problems with securitization and trading of receivables from *murabaha* and *tawarruq* facilities. Those problems arose from the fact that most jurists, with the notable exception of those in Malaysia, forbade trading of debts, except under very strict transfers at face values and resale to the debtor. This prevented the development of secondary markets that would allow banks to diversify their portfolios and sources of funds. Lease financing provided a partial solution to both problems: it was ostensibly based on real assets that continued to play a role throughout the life of the financial facility, and it was possible to trade lease receivables on secondary markets as ostensible shares in the leased assets.

Jurists were adamant that Islamic lease or *ijara* financing must be truly asset-based, and therefore must be structured as operating rather than financial leases. However, recent advances in structured finance – which helped corporations such as Enron to move debts and interest payments off balance sheets through sale-leaseback structures – had blurred the line between operating and financial leases. As a result, a prestigious juristic council declared in 2008 that more than 80 per cent of lease-based bond (*sukuk*) structures were unIslamic, since material ownership of the underlying assets was not real.

Developed initially as another mode of secured lending, lease financing proceeded by acquiring durable assets and leasing them with an option to buy – principal plus interest passing to the lessor as rent plus potential final payment. For banks in countries that forbid them from owning real estate, special purpose vehicles (SPVs) received credit that were used to acquire the assets and lease them till maturity. Shares in those SPVs were treated as shares in the leased properties, thus allowing them to trade on secondary markets. In the United States, such structures were used to originate mortgage loans that were then securitized through Fannie Mae and Freddie Mac, and marketed both domestically and in the cash-rich GCC, especially after the second wave of petrodollar flows began in 2001.

Bond structures were easily adapted from these financial forms. An entity that wished to issue a bond would create an SPV, which sold shares for the amount of financing desired. The proceeds of that sale were used to buy some asset from the originator, which asset was promptly leased back. The originator would thus collect the proceeds of the sale of its asset as principal, and pay principal plus interest in the form of rent and/or a final repurchase price, which payments were passed through to the *sukuk* or bond holders. An added advantage of this structure is that the payments were made ostensibly on shares in ownership of the real asset, thus the contract could be advertised as a form of partnership, which appealed to the earlier political-Islam inspired literature on Islamic economics.

Jurists further facilitated securitization of debts by allowing a portfolio of asset-based and purely debt-based receivables (that is, lease-based and credit-sale-based, respectively) to be traded as long as the asset-based component exceeded 51 per cent of the total face value (Usmani 1998). This strange provision clearly imposed no significant constraints on securitization, since successive portions of pure-debt receivables could be bundled iteratively with the same asset-based ones, which could be bought back repeatedly for the purpose of bundling with pure-debt tranches. Thus, Islamic finance became an equal partner in the credit bubble the ensued in the first decade of the 21st century. In fact, the volume of *sukuk* remained sufficiently small (relative to demand by Islamic banks) to merit abnormally high prices and low yields relative to conventional debts issued by the same entities.

Islamic Mutual Funds

A widely publicised area of Islamic finance was the development of 'screening' methods to identify '*Shari*'*a* compliant' stocks. These screens excluded stocks of companies with significant forbidden activities (such as breweries), and also of firms with excessive debt or interest income. The debt screen chosen by the industry was particularly perplexing, as it excluded firms with debt to market capitalization ratios exceeding one-third. This rule clearly forced fund managers to buy high and sell low in highly volatile markets. Moreover, the rule diverted funds away from Muslimowned companies, which were not allowed any degree of unsecured-loan leverage, in favour of western firms with moderate levels of leverage. The financial screens themselves had no foundation in Islamic law or reasonable economic analysis, starting as they did at 5 per cent debt to assets and evolving during the tech-stock bubble of the late 1990s into 33 per cent of debt to assets and then 33 per cent of debt to market capitalization. It is not clear whether and when these rules can be replaced with sensible ones.

Takaful (Islamic Insurance) and Derivatives

One of the fast-growing sectors in Islamic finance is an Islamic alternative to commercial insurance known as *takaful* (mutual support). The rhetoric of this sector is based on the idea of mutual protection against losses, but most *takaful* companies to date have not been structured as mutual insurance companies (in which policyholders and shareholders are the same individuals). Instead, *takaful* companies are generally shareholder-owned and act through silent partnership or agency to invest the policyholders' premiums and pay legitimate claims in the form of 'voluntary contributions' – thus avoiding the Islamic prohibition of *gharar*, which includes trading known amounts (policy premia) for uncertain future amounts (on potential valid insurance claims). The prohibition of *gharar* was also invoked to forbid derivative securities, but forwards and options were easily synthesised from the ancient contracts of *salam* (prepaid forward sale) and '*urbun* (downpayment call option), respectively.

Substance and Form

El-Gamal (2006, 2008) has argued that the essence of the ancient religious law was

regulatory. It is well known in financial economics that financial innovators eventually find means to circumvent outdated regulation, thus increasing systemic risk. Financial crises later propel political and economic authorities to impose further regulations for innovators to circumvent. In this regard, the ancient religious regulations enshrined in medieval Islamic jurisprudence, especially if interpreted naively as prohibitions of certain contracts and permissions of others, are woefully out of date, and therefore ceased to perform their regulatory function centuries ago. Indeed, that is precisely why majority-Muslim societies had abandoned those outdated contract-based frameworks before the Islamist revisionism of the mid-20th century. The ancient law, which is not uniquely Islamic, does contain many lessons for today's societies – Muslim and otherwise. However, rent-seeking behaviour by bankers, lawyers and religious scholars on the one hand, and incoherent pietism and adherence to fictional Utopian history on the other, have prevented societies from adapting this centuries-old accumulated human wisdom for any purpose beyond short-term self-enrichment and identity-political appeasement, both of which increase rather than ameliorate systemic risks.

See Also

► Islamic Economic Institutions

Bibliography

Chapra, U. 1996. *What is Islamic economics?* IDB prize winner's lecture series, vol. 9. Jeddah: Islamic Development Bank.

El-Gamal, M. 2003. Interest and the paradox of contemporary Islamic law and finance. *Fordham International Law Journal* 27(1): 108–149.

El-Gamal, M. 2006. *Islamic finance: Law, economics, and practice*. Cambridge: Cambridge University Press.

El-Gamal, M. 2008. Incoherence of contract-based Islamic financial jurisprudence in the age of financial engineering. *Wisconsin International Law Journal* 25(4): 605–623.

Haneef, M. 1995. *Contemporary Islamic economic thought: A selected comparative analysis*. Kuala Lumpur: Iqrak.

Humud, S. 1976. *Tatwir Al-A'mal Al-Masrifiyya bima Yattafiqu wa Al-Shari'ah Al-Islamiyya* (Evolving banking transactions in accordance with Islamic law). Cairo: Dar Al-Ittihad Al-'Arabi lil-Tiba'a.

Islamic Financial Services Board (IFSB). 2006. *Guiding principles on corporate governance for institutions offering only Islamic financial services…*. Kuala Lumpur: IFSB. Available at http://www.ifsb.org/standard/ifsb3.pdf. Accessed 12 Jan 2009.

Kuran, T. 2004. *Islam and Mammon*. Princeton: Princeton University Press.

Reisman, Y. 1995. *The laws of ribbis*. New York: Mesorah Publications.

Siddiqi, M.N. 1983. *Banking without interest*. Leicester: Islamic Foundation.

Tantawi, M. 2001. *Mu'amalat Al-Bunuk wa Ahkamuha Al-Shar'iyya*. Cairo: Nahdat Misr.

Udovitch, A. 1970. *Partnership and profit in medieval Islam*. Princeton: Princeton University Press.

Usmani, M.T. 1998. *An introduction to Islamic finance*. Karachi: Idaratul Ma'arif.

IS–LM

Steven N. Durlauf and Donald D. Hester

Abstract

The IS–LM model is a short-run macroeconomic analytical construct for studying an economy with idle productive resources. The diagram has been especially influential because its constituent curves are loci on which the goods market (IS curve) and the money market (LM curve) are respectively in equilibrium, making it possible to infer changes in fiscal policy and monetary policy, both separate and simultaneous. The model is prominent in elementary and intermediate macroeconomic textbooks, yet it fails to accommodate the main features of modern macroeconomic theory, although modern dynamic models are sometimes interpreted as having IS–LM type features.

Keywords

Asset price equilibrium; Bretton Woods System; Central bank independence; Crowding out; Deflation; Econometric Society;

Expectations; Federal Reserve System; Fiscal drag; Fiscal policy; Fleming, J.; Floating exchange rate; General equilibrium; Government budget constraint; Hicks, J.; Inflation; Inflationary expectations; International capital markets; IS–LM model; Kahn, R.; Keynes, J. M.; Liquidity trap; Lump sum taxes; Macroeconomic volatility; Marshall–Lerner condition; Microfoundations; Monetarism; Monetary policy; Monetary policy rules; Money demand; Money supply; Mundell, R.; Mundell–Tobin effect; Open market operations; Pigou effect; Price control; Progressive taxation; Real balance effect; Robinson, J.; Stabilization policy; Sticky prices; tâtonnement; Taxation of income; Taylor rule; Uncertainty; Walras's Law

JEL Classifications
E1

The IS–LM model is a short-run macroeconomic analytical construct for studying an economy with idle productive resources. In the form exposed by Hansen (1949), it is a two-dimensional diagram with the abscissa measuring real income and the ordinate the real interest rate. It has been widely and successfully employed in interpreting macroeconomic policy and is prominent in elementary and intermediate macroeconomic textbooks. A close antecedent, the SI–LL diagram, first appeared in print in an influential article by J.R. Hicks (1937) that proposed an interpretation of Keynes's *General Theory* (1936) and effectively made Keynes's contribution accessible to large numbers of students ever since. Hicks's diagram had nominal income measured on the abscissa and was unclear about whether the interest rate was real or nominal. If prices are fixed or 'sticky' these differences are immaterial, but assumptions about prices and their measurement loomed large in subsequent controversies and applications of the diagram. Lange (1938) appears first to have required that variables were real magnitudes. Reflecting the times of its origin, the model describes a closed economy.

The diagram has been especially influential because its constituent curves are loci on which the goods market (IS curve) and the money market (LM curve) are respectively in equilibrium. The intersection of the two curves is a point where both markets (and, through Walras's Law, the bond market) are in equilibrium. The labour market is not required to be in equilibrium. Because fiscal policy affects the goods market through tax, transfer and expenditure changes, the effects of fiscal policy can be inferred from the change in the intersection of the IS curve with a stationary LM curve. Similarly, because changes in the money stock affect only the LM curve, the effects of monetary policy can be inferred from the change in the intersection of the LM curve with a stationary IS curve. Finally, the effects of simultaneous changes in both fiscal and monetary policies can be predicted from the change in the intersection when both curves are moved.

By way of background, the *General Theory* contains no formal mathematical model and has only one diagram. 'Keynes believed economics was over-addicted to "specious precision" – making perfectly precise what was in reality vague and *complex*. It is significant that he refused to present the "model" of the *General Theory* in mathematical form, even though he assembled its (verbal) elements in chapter 18' (Skidelsky 1994, p. 540). 'The mathematicisation of the *General Theory* started immediately it (*sic*) was published but it was left to Hicks to map the mathematics on to a two-curve diagram which became the accepted form of the *General Theory*' (Skidelsky 1994, p. 611).

Hicks's article emerged from the September 1936 European meetings of the Econometric Society at Oxford where a symposium on 'Mr. Keynes' System' was held. Other important papers from the symposium interpreting Keynes were by R.F. Harrod (1937) and J. Meade (1937); both were published slightly earlier than Hicks's paper, but contained no trail-blazing graphical apparatus. Young (1987, p. 29) claims that all three papers had the same underlying equation system, which differed from that of the *General Theory* but may have appeared in Keynes's lectures at Cambridge as early as 1934. All three papers analysed the relation between their interpretation of what underlay the

General Theory and the pre-existing theoretical framework. Keynes apparently did not object to the specifications of Hicks, Harrod and Meade, but stressed the importance of expectations and uncertainty in his subsequent discussions (1937) of the general theory; expectations do not appear formally in the three authors' equation systems. If one accepts Keynes's (1936, ch. 12) discussion of how long-term expectations are formed, it may indeed be specious to describe commodity market equilibrium as if it were lying on a stationary curve, but that in no way reduces the usefulness of the model for designing and interpreting policy. The effects of monetary and fiscal policy actions are unaffected by random shocks to the two curves. However, the effects might be affected if the curves are moved by expectations about present or future policy moves, as has been suggested by Lucas (1976). Subsequently, J. Robinson (1975) and R. Kahn (1984), two of Keynes's contemporaries when the *General Theory* was being drafted, objected strenuously to the IS–LM formulation and Hicks himself (1982) indicated dissatisfaction with it.

Basic Theoretical Structure

In the standard formulation of the IS–LM model, the endogenous variables to be determined are the level of aggregate output Y and the real interest rate *r*. Aggregate demand in the goods market is modeled via the output identity:

$$Y = C + I + G + NX$$

where C denotes consumption, I investment, G government spending and NX net exports. The identity is given substance by replacing C with a consumption function and I with an investment function. Typically, IS–LM analysis assumes that consumption depends positively on disposable income, which equals output Y minus taxes $T(Y)$ (income taxes induce dependence of the level of taxes that are collected on income) whereas investment depends on the real interest rate; one could also endogenize government spending and net exports. This leads to the IS equation:

$$\text{IS}: Y = C(Y - T(Y)) + I(r) + G + NX.$$

Money market equilibrium is defined by equating money demand and money supply. Real money demand, denoted by L to capture the idea that the demand for money is the demand for liquidity, is assumed to depend negatively on the nominal interest rate, which by definition equals the real interest rate plus the expected inflation rate, π. Throughout, expected inflation will be treated as exogenous; as noted below a defect of the IS–LM framework is that it does not embody expectations in an interesting way. The real money supply, $\frac{M}{P}$, is treated as exogenous, as are the price level and inflation rate. The LM equation is:

$$\text{LM}: \frac{M}{P} = L(r + \pi, Y).$$

If the demand for money does not depend on the nominal interest rate, then the LM curve uniquely determines the level of output. This special case is of historical importance in understanding the monetarist perspective on macroeconomics. In contrast, if the demand for money is infinitely elastic at some exogenous nominal interest rate, the LM curve uniquely determines the real interest rate. This case is also of historic interest as it is the first version of a liquidity trap. It should also be noted that the LM curve can be replaced with a more sophisticated system of asset price equilibrium conditions; a significant component of James Tobin's work on monetary economics well summarized in Tobin (1969), represented an effort to enrich IS–LM analysis via a richer specification of financial markets.

Regardless of the specific assumptions on the shapes of the IS and LM schedules, the equilibrium pair (r, Y) is determined by the simultaneous solution of these two equations. Comparative static analysis may be done by changing the various exogenous variables in the IS and LM equations. Notice that this system is entirely demand driven in the sense that it does not consider resource constraints in the determination of output.

The IS–LM model has a number of well-known implications with respect to the effects of changes in government policy. Increases in government

spending G increase the equilibrium levels of r and Y and decrease the equilibrium level of I. The reduction in investment induced by an increase in government spending is known as crowding out. When money demand does not depend on the nominal interest rate, there is complete crowding out in the sense that the increase in G is completely offset by a decrease in I, so that aggregate demand is unaffected; the independence of money demand from the nominal interest rate has often been treated as a hallmark of monetarism, since in this case changes in fiscal policy have no real effects. Similar results occur when one considers tax changes; a reduction in either lump sum taxes or the income tax rate increases both Y and r. With respect to monetary policy, an increase in M leads to an increase in Y, a decrease in r and an increase in I. Hence, unlike expansionary fiscal policy, expansionary monetary policy increases investment, and so causes crowding in. The exception to this result is a liquidity trap in which changes in the supply of money are accepted by the public at the initial nominal interest rate.

Beyond the evaluation of exogenous changes in government policies on aggregate outcomes, the IS–LM model has also been used to evaluate alternative government policies. Poole (1970) is particularly notable in this regard. Poole compares the stabilization properties of a monetary policy that fixes the nominal money supply with one that fixes the real interest rate. He shows that, if macroeconomic volatility derives from shocks to the IS schedule, then a fixed money stock policy stabilizes an economy more than a fixed interest rate policy; in contrast, when aggregate fluctuations are generated solely by shocks to the LM schedule, a fixed interest rate completely eliminates aggregate fluctuations. The idea that the effects of a monetary policy rule depend on the type of shocks an economy experiences has proven to be of importance in contexts far beyond the IS–LM model; one sees echoes of Poole's reasoning in discussions of the Taylor rule for interest-rate setting.

The IS–LM model has also been used to study the effects of changes in other exogenous (from the perspective of the model) variables on the macroeconomic equilibrium. An especially important issue concerns changes in the price level, because one wants to know whether price adjustments can move aggregate output towards a level consistent with full employment. In the specification so far described, a decrease in the price level raises Y and lowers r because a lower price level increases the real money supply, thereby shifting the LM schedule. This property was considered important in early expositions of IS–LM such as Modigliani (1944) that considered the possibility of a liquidity trap.

However, as argued by Pigou (1943, 1947), there is another channel through which lower prices can raise demand. Pigou argued that the level of consumption depends on the level of wealth as well as on disposable income. Since a component of wealth is nominal money, price reductions can increase demand through increases in the real value of money and, hence, wealth. This is called the Pigou (or real balance) effect. Work on the real balance effect, in turn, affected the study of monetary policy in the IS–LM framework. The seminal paper in this regard is Metzler (1951), who argued that the real balance effect implied an important difference between the effects of a helicopter dropping of money, that is, an increase in the money supply in which additional money are simply added to individual portfolios, and an open market operation, in which an increase in the money supply is generated by the trading of money for bonds, so that one nominal asset is swapped for another, thereby keeping the aggregate nominal supply of assets constant.

Another well-known property of the model concerns the equilibrium effects of an increase in the (exogenously given) inflation rate. In the IS–LM model, an increase in π increases Y and lowers r. Intuitively, an increase in inflation reduces money demand, requiring an adjustment of the real interest rate and output to compensate. The fact that increases in inflation lead to less than one-to-one increases in the nominal interest is known as the Mundell–Tobin effect (Mundell 1963a; Tobin 1965). As before, this property depends on the dependence of money demand on the nominal interest rate.

From the perspective of modern macroeconomic theory, the IS–LM model has very serious deficiencies. One problem is that the model lacks

I

well-defined microeconomic foundations. The price level is treated as exogenous; while prices may be sticky, the complete rigidity found in the IS–LM model is unappealing. Further, the consumption and investment functions are typically specified in an ad hoc fashion, rather than as the outcome of solving explicit decision problems.

Also, the IS–LM model fails to embody aggregate dynamics. The model constructs a snapshot of the macroeconomy without accounting for the fact that the snapshot is really one frame of a motion picture. While expectations variables can be introduced into the model, its static nature precludes one from considering many implications of the intertemporal government budget constraint for the real effects of changes in fiscal policy or the effects of expectations about monetary policy on the sequence of equilibrium price levels over time. This lack of dynamics was recognized early on as a defect of the model; heuristic analyses include Patinkin (1956) who tried to link IS–LM with tâtonnement adjustment of prices à la Walras. Indeed, Patinkin's book was the high point of efforts to link IS–LM with the conceptual structure of general equilibrium models. But this type of work disappeared rather quickly. Later on, efforts were made by Blinder and Solow (1973) and Tobin and Buiter (1976) to account for changes in the stocks of various assets on the IS–LM equilibrium, with particular attention paid to understanding how permanent changes in fiscal policy affect output in the presence of the requirement that the government budget balance with respect to the present discounted value of debt and taxes. These analyses found that accounting for such effects could imply that the long-run effect of a change in government spending exceeds the short-run change. One reason for this is the increased holdings of government debt induced by a fiscal expansion will increase consumption. This type of work also failed to have much effect on the use of the IS–LM framework.

To be fair, a number of recent authors have attempted to provide more rigorous microeconomic foundations to the IS–LM model; McCallum and Nelson (1999) is the most important example. See King (2000) for evaluation of such models. While progress has been made in producing better microeconomics for the IS–LM model, it seems fair to say that much of its use, especially in pedagogical and policy contexts, relies on the model we have described.

Mundell–Fleming Model

A variant of the IS–LM model that has proven important in international economics is due to Mundell (1963b) and Fleming (1962); its value derives from the consideration of how the effects of monetary and fiscal policy are altered when one considers the role of the exchange rate. In this framework, a small country is assumed so that the income of the rest of the world, Y^{ROW}, is unaffected by events in the country. However, actions of the country may affect the exchange rate e, defined as the number of units of foreign currency one unit of the country's currency can purchase. Net exports for the country are assumed to obey

$$NX = NX\left(Y, Y^{ROW}, e\right).$$

Higher exchange rates are assumed to lead to a lower level of net exports; this is known as the Marshall–Lerner condition.

The effects of changes in monetary and fiscal policy will critically depend on the effects of a policy on the exchange rate. This, in turn, depends on the degree of integration of international capital markets. Following the classic Mundell–Fleming analysis, suppose that international capital markets are fully integrated. In this case, the real return on investments cannot differ across countries and so r is fixed since the economy under study is small. An increase in government spending by a small economy, in this case, will have no real effects. An increase in G will be entirely offset by exchange rate appreciation, that is, an increase in e, so that there is no net fiscal stimulus as the increase in government spending is fully offset by a reduction in net exports. In contrast, an increase in the money supply will induce an increase in output via exchange rate depreciation. These outcomes, of course, presume that the exchange rate is allowed to float.

In contrast, suppose that the central bank authority is committed to maintaining a given exchange rate, $\bar{e}$. In this case, the effects of monetary and fiscal policy are quite different. Maintenance of the exchange rate eliminates any independent role for the central bank in the sense that any action it takes to raise output will have to be undone in order to preserve the exchange rate. In contrast, a fiscal stimulus will induce a subsequent increase in the money supply in order to overcome the associated exchange rate appreciation, which reinforces the effects of the stimulus that are found in the closed economy model.

While current thinking on the interactions of the exchange rate regime with policy effects has moved far beyond the details of the Mundell–Fleming model, the ideas in the model not only proved of direct value for much subsequent research (for example, Tobin and Braga de Macedo 1980) but has also, via its limitations, defined the agendas of alternative research directions (Obstfeld 2001).

IS–LM in the Light of Macroeconomic History

Despite its theoretical limitations, the IS–LM model is illuminating in interpreting macroeconomic policy and events in the post-war period, especially in the United States, which could reasonably be viewed as a closed economy in the early years. Between the end of the Second World War and 1951, the Federal Reserve was committed to a policy of restricting upward shifts in the yield curve in order to reduce the cost of financing the US government's large war debt. During this period, the Federal Reserve effectively 'pegged' the short-term nominal interest rate, so that monetary policy was expansionary whenever the real rate was negative and especially so when inflation was increasing. In the inflationary period between 1945 and 1948, the real short-term interest rate was substantially negative. Although deflation accompanied the recession of 1949, when the real rate turned briefly positive this unsustainable pegging policy together with the onset of the Korean War in 1950 led to the 'Accord' of 4 March 1951, which permitted the Federal Reserve to undertake discretionary monetary policy. This early inflationary monetary policy bias may not have been inappropriate, because demobilization after the Second World War led to a massive leftward shift of the IS curve due to falling government spending and essentially stagnant real net foreign and gross domestic investment until 1950. Inflation occurred with the suspension of price controls in 1946, but did not begin to accelerate until the war started in June 1950.

The Korean War led to a 75 per cent increase in real government spending and a sizable increase in the real government deficit between 1950 and 1954. An expansion of accelerated depreciation allowances in 1954 was associated with a ten per cent increase in real investment in producers' durable equipment in the subsequent three years. Both events caused a rightward shift in the IS curve. However, the shifts were rather insidiously being offset by 'fiscal drag' that resulted from increases in the marginal income tax rate that existed when the economy was at full employment. The tax rate rose because progressive tax schedules applied to nominal income, not income adjusted for inflation. Following the Accord, the Federal Reserve used open-market operations to fight inflation by raising nominal interest rates on several occasions with unsatisfactory results. Unemployment rates rose as interest rates rose, as might have been predicted from the resulting leftward shift in the LM curve. However, it was not widely understood that the IS curve was also shifting leftward because of the cumulative effects of fiscal drag. Seemingly, restrictive monetary policy induced three recessions in this decade with unemployment rates at troughs successively higher in each recession. Inflation temporarily abated during or shortly after each recession, but then returned, in part, because real short-term interest rates were infrequently positive until 1959.

The 1960 elections resulted in John F. Kennedy becoming president and, of greater significance for this discussion, a generation of economic advisors who understood and were intent upon applying the IS–LM model. The US Council of Economic Advisors' *Economic Report of the President* explicitly

focused on the importance of the full-employment budget surplus (1962, pp. 78–81) and, thus, fiscal drag. The Council of Economic Advisors also worried about the fact that real gross private domestic investment had been below its 1955 peak for the subsequent six years, a possible consequence of high interest rates. The administration negotiated an arrangement with the Federal Reserve whereby it would attempt to twist the yield curve through open-market operations by increasing short-term interest rates to protect the US gold stock from foreign withdrawals and lowering long-term interest rates to stimulate investment. The Treasury assisted in this effort with its debt management policies (1962, pp. 86–91). The administration increased federal government spending significantly beginning in the 1962 fiscal year and would subsequently stimulate the economy with tax cuts. The federal government deficit rose over the four years after 1960 and the IS curve shifted rightward. Because twisting of the yield curve involved two interest rates, it cannot be interpreted directly with the LM curve; however, between 1961 and 1964 the interest rate on three-month treasury bills rose 50 per cent, the rate on three–five year issues rose 6.4 per cent, and between December 1960 and December 1964 the level of Federal Reserve credit outstanding rose by 37 per cent. In part because of continuing gold outflows, the last was the largest rate of growth of Federal Reserve credit until that date over any four-year span since the Accord. Net of gold flows Federal Reserve credit expanded by 28 per cent in this period of relatively low inflation. It is hard to argue that the LM curve didn't also shift rightward.

Tax cuts were phased in 1962 in the form of an investment tax credit that effectively reduced the required rate of return that profitable firms needed to undertake an investment project and in 1964 and 1965 in the form of 10 per cent reductions in corporate and personal income tax rates. The investment tax credit and accommodating monetary policy was associated with a 50 per cent increase in real gross private domestic investment between 1961 and 1966; the tax credit implies a rightward shift in the IS curve. The large tax rate cuts in 1964 and 1965 did not lead to an increase in the federal government deficit, partly because GDP rose considerably in response to rising investment. The unemployment rate fell from 6.7 per cent in 1961 to 3.8 per cent in 1966 and the average annual rate of inflation from the end of 1960 to the end of 1966 was less than two per cent, although it began to rise in the fourth quarter of 1965.

Because of rising inflation, a policy change occurred at the end of 1965 when the Federal Reserve signalled with an increase in its discount rate that it would begin to restrict credit. All interest rates (real and nominal) rose sharply in 1966 and the real money supply fell for four successive quarters; the LM curve was shifting leftward. The Federal Reserve was briefly successful in reducing the rate of increase in prices at the end of 1966, but then inflation rose sharply in 1967 and 1968 as large deficits resulted from the Vietnam War. One reason inflation rose was that the Federal Reserve was focusing on nominal rather than real short-term interest rates; the latter were negative on average in the last three quarters of 1967 and so monetary policy was actually expansionary. The IS curve was shifting rightward until a temporary ten per cent income tax surcharge was imposed in January 1968 on corporate income taxes and in April 1968 on personal income taxes. The federal budget deficit in the national income accounts turned into a surplus in the third quarter of 1968. Beginning in early 1968, the Federal Reserve began to raise real interest rates dramatically. With both the IS and LM curves shifting leftward, the economy began to slow and the unemployment rate began to rise in 1969, as the model predicts.

During the 1960s it was becoming less tenable to view the United States as a closed economy. Although the percentage of US exports to GNP had risen from six per cent in 1946 only to seven per cent in 1969, international events were beginning to impair the usefulness of the original Hicksian model. The quasi-fixed exchange rate system that had been established in the 1944 Bretton Woods Conference began to collapse in 1968 when the US gold stock reached a critically low level. In light of the aforementioned important contributions by Mundell and Fleming, which argued that monetary policy is ineffective and fiscal policy is powerful in a fixed-exchange regime with perfect capital markets, it is necessary

to digress to explain how the model applied before the actual collapse in 1971–3 and afterwards.

At the end of the Second World War many countries restricted currency and capital flows, which allowed both fiscal and monetary policy to be effective, as in the world Hicks envisioned. These restrictions were gradually relaxed during the subsequent years. As they disappeared, the efficacy of monetary policy weakened, although imperfect capital markets allowed it to have some residual potency. Monetary policy was weakened because countries were obliged to maintain quasi-fixed exchange rates by not allowing real interest rates to vary across countries. In contrast, fiscal policy was strengthened because central banks were obligated to take actions that offset the effects of fiscal actions on real interest rates, essentially causing the LM curve to shift in the same direction that the IS curve shifted.

By 1973 the world was in a 'dirty' floating exchange rate system where various countries attempted to maintain some fixed bilateral exchange rates with their major trading partners. Mundell and Fleming had argued that in a pure floating exchange rate system with perfect capital mobility fiscal policy would be ineffective and monetary policy would be very strong, because any action to change a country's real interest rate relative to other countries would be reinforced by a change in its trade balance. In other words, a shift in its LM curve would be reinforced by a shift in its IS curve in the same direction because its trade balance was negatively related to its exchange rate, which was positively related to the value of its real interest rate relative to those of its trading partners. Because countries were unwilling to have their exchange rates be completely flexible, fiscal policy was considerably weakened relative to the fixed exchange rate period but still continued to have some power. Monetary policy was strengthened.

This open economy extension of the IS–LM model has proven to be illuminating about monetary and fiscal policy in the post-1971 period, again particularly in the United States. With the change in the exchange-rate regime, the trade-weighted value of the dollar fell about 20 per cent between 1971 and 1973, which, together with a recession in 1974, was sufficient to allow the United States to have a trade surplus on average through 1976. Between 1973 and 1979 the Federal Reserve allowed the real short-term (federal funds) rate to be negative on average, which led to substantial inflation and a bubble in the housing market. However, the international trade-weighted value of the dollar was essentially unchanged between the middle of 1973 and the middle of 1978, because average nominal short-term interest rates and inflation in major trading partners of the United States moved in tandem. Beginning in 1977 the US trade deficit began to increase and after mid-1978 the value of the US dollar fell unevenly until July 1980.

In July 1980 the Federal Reserve began to reduce the real money stock and, with accompanying large tax cuts in 1981–3, real and nominal interest rates rose to record levels, actions that were not offset by matching policies in foreign countries. As a result, the trade-weighted dollar appreciated from 84.65 (March 1973 = 100) in July 1980 to 158.43 in February 1985 and the trade deficit soared. The IS curve shifted to the right because the increase in the federal deficit was larger than the increase in the trade deficit during these years; the LM curve shifted to the left. As the IS–LM model predicts, the expansionary effects of the 1981–3 tax cuts, as measured by changes in real GDP and the unemployment rate, were much smaller than those of the similarly sized 1964–5 tax cuts, because of both the non-accommodating monetary policy and the dollar's appreciation.

In September 1985 a meeting of representatives of five major nations in New York resulted in a successful coordinated effort to reduce the trade-weighted value of the dollar, which fell 30 per cent in the succeeding two years and was followed by a sharp reduction in the US trade deficit. US short-term real interest rates fell until the middle of 1988 and the unemployment rate reached a low of 5.2 per cent in 1989. As the extended IS–LM model predicts, monetary policy was quite effective. Monetary policy effectiveness would be repeatedly evident in the following years.For example, the Federal Reserve raised real short-term interest rates between July 1988 and December 1990 to combat inflation, which

resulted in a short recession in 1991. It successfully responded to rising unemployment by cutting its real overnight federal funds interest rate to near zero in 1993, which dramatically lowered the unemployment rate in 1995. By sharply raising this interest rate in 1994, the Federal Reserve managed to continue to lower the unemployment rate with negligible inflation until a stock-market bubble burst in 2001. Both fiscal and monetary policies were strongly expansionary between 2001 and 2005. The real federal funds rate had on average been negative since the end of 2001. As might have been predicted from the Mundell–Fleming model, the trade deficit expanded enormously; its increase was roughly equal to the increase in the federal government deficit.

Finally, a troubling problem with prolonged periods of negative short-term interest rates is that they have tended to manifest themselves in rapid rates of inflation in prices of houses – both in the 1970s and in the 2000s. As Keynes and his followers warned, expectations are at the heart of the *General Theory* and they are not prominent in the textbook expositions of the IS–LM model. Many macroeconometric models can be interpreted as extensions of the textbook IS–LM model, and as time has passed many of them have increasingly attempted to incorporate expectations formation. Expectations are prominent in the Federal Reserve's recent model (Brayton et al. 1997), but their formation may not yet be accurately represented.

Conclusions

The IS–LM model occupies an awkward position in modern macroeconomics. It is still a workhorse of undergraduate teaching and still widely used by economists in developing intuition about short-run macroeconomic phenomena, including policy counterfactuals; see Colander (2004) for discussion of these roles. However, the model fails to accommodate the main features of modern macroeconomic theory, although modern dynamic models are sometimes interpreted as having IS–LM type features. We expect this dichotomy and this anomalous use to continue.

See Also

► Monetarism

Bibliography

Blinder, A., and R. Solow. 1973. Does fiscal policy matter? *Journal of Public Economics* 2: 319–337.

Brayton, F., E. Mauskopf, D. Reifschneider, P. Tinsley, and J. Williams. 1997. The role of expectations in the FRB/US macroeconomic model. *Federal Reserve Bulletin* 83(4): 227–245.

Colander, D. 2004. The strange persistence of the IS–LM model. *History of Political Economy* 36(Suppl. 1): 305–322.

Fleming, J. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9: 369–379.

Hansen, A. 1949. *Monetary theory and fiscal policy*. New York: McGraw-Hill Publishing Co..

Harrod, R. 1937. Mr. Keynes and traditional theory. *Econometrica* 5: 74–86.

Hicks, J. 1937. Mr. Keynes and the 'classics': A suggested interpretation. *Econometrica* 5: 147–159.

Hicks, J. 1982. IS–LM: An explanation. In *Money, interest, and wages: Collected essays on economic theory*, Vol. 2. Oxford: Basil Blackwell.

Kahn, R. 1984. *The making of Keynes's general theory*. Cambridge: Cambridge University Press.

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

Keynes, J.M. 1937. The general theory of employment. *Quarterly Journal of Economics* 51: 209–223.

King, R. 2000. The new IS–LM model: Language, logic and limits. *Federal Reserve Bank of Richmond Economic Quarterly* 86(3): 45–103.

Lange, O. 1938. The rate of interest and the optimum propensity to consume. *Economica N.S.* 5: 12–32.

Lucas, R. Jr. 1976. Econometric policy evaluation: A critique. In *The phillips curve and labor markets*, *Carnegie Rochester conference series on public policy*, ed. K. Brunner and A. Meltzer, Vol. 1. Amsterdam: North-Holland.

McCallum, B., and E. Nelson. 1999. An optimizing IS–LM specification for monetary policy and business cycle analysis. *Journal of Money, Credit & Banking* 31: 296–316.

Meade, J. 1937. A simplified model of Mr. Keynes' system. *Review of Economic Studies* 4(2): 98–107.

Metzler, L. 1951. Wealth, saving, and the rate of interest. *Journal of Political Economy* 59: 93–116.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12: 45–88.

Mundell, R. 1963a. Inflation and real interest. *Journal of Political Economy* 73: 280–283.

Mundell, R. 1963b. Capital mobility and stabilization policy under fixed and flexible exchange rates. *Canadian*

Journal of Economics and Political Science 29: 475–485.
Obstfeld, M. 2001. International macroeconomics: Beyond the Mundell–Fleming model. *IMF Staff Papers* 47: 1–38.
Patinkin, D. 1956. *Money, interest, and prices*. Evanston: Row Peterson.
Phelps Brown, E. 1937. Report of the Oxford meeting, September 25–29, 1936. *Econometrica* 5: 361–383.
Pigou, A. 1943. The classical stationary state. *Economic Journal* 53: 343–351.
Pigou, A. 1947. Economic progress in a stable environment. *Economica* 14: 180–188.
Poole, W. 1970. Optimal choice of optimal monetary instruments in a simple stochastic macro model. *Quarterly Journal of Economics* 84(2): 197–216.
Robinson, J. 1975. What has become of the Keynesian revolution? In *Essays on John Maynard Keynes*, ed. M. Keynes. Cambridge: Cambridge University Press.
Skidelsky, R. 1994. *John Maynard Keynes: The economist as savior 1920–1937*. New York: Viking Penguin.
Tobin, J. 1965. Money and economic growth. *Econometrica* 33: 671–684.
Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1(1): 15–29.
Tobin, J., and J. Braga de Macedo. 1980. The short-run macroeconomics of floating exchange rates: An exposition. In *Flexible exchange rates and the balance of payments: Essays in memory of Egon Sohmen*, ed. J. Chipman and C. Kindleberger. Amsterdam: North-Holland.
Tobin, J., and W. Buiter. 1976. Long run effects of fiscal and monetary policy on aggregate demand. In *Monetarism*, ed. J. Stein. Amsterdam: North-Holland.
US Council of Economic Advisors. 1962. *Economic report of the President together with the annual report of the Council of Economic Advisors*. Washington, DC: US Government Printing Office.
Young, W. 1987. *Interpreting Mr Keynes: The IS–LM enigma*. Cambridge: Basil Blackwell.

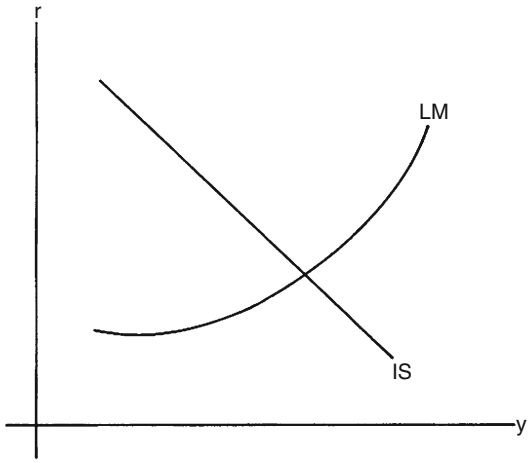
IS–LM Analysis

Axel Leijonhufvud

The original IS–LM model was introduced by Sir John Hicks as a framework for clarifying the relationship between Keynes's theory and that of his predecessors. (In Hicks's famous paper, 'Mr Keynes and the "Classics"' (1937), however, the now so familiar diagram bore the notation SI-LL.) Further attempts to define Keynes's theoretical contributions precisely within the basic IS–LM structure were made by Alvin H. Hansen (e.g. Hansen 1953), Franco Modigliani (1944), Lawrence Klein (1947) and Don Patinkin (1948) among others. IS–LM became in this way not only the vehicle for popularizing Keynesian ideas and the mainstay of macroeconomic textbooks but, for several decades, the main organizing conception for macroeconomics in general. Even the very large macroeconometric models of several hundred equations were generally disaggregated IS–LM structures. When Modigliani (1963) surveyed the major developments in macroeconomics in the early 1960s, he did so by presenting an 'updated' IS–LM model. As late as 1971, Milton Friedman and his critics debated the issues between Monetarism and Keynesianism in accordance with IS–LM groundrules (cf. Gordon 1973). From the late 1970s on, the grip of IS–LM on the macrotheoretical imagination began to loosen as the rational expectations movement came to rely on smallscale general equilibrium and game-theoretic models instead.

The equations of the basic model come in three blocks. The IS block, in the simplest version, consists of a consumption-(or savings-) function, an investment function, and a saving-equals-investment equilibrium condition. Government expenditures and taxation are optional features. The LM block, consists of a money demand (or liquidity preference) equation, a money supply equation, and an equilibrium condition for money. The employment block has an aggregate production function, from which is derived a labour demand function: the unemployment version of the model is usually completed by adding the restriction that the money wage cannot fall below a certain specified value ('rigid wages'); the full employment version has instead a labour supply function and an equilibrium condition for the labour market.

Each of the first two blocks can be reduced to a single relationship between income and the interest rate. The two reduced forms in turn produce the familiar diagram shown in Fig. 1.



IS–LM Analysis, Fig. 1 IS-LM

The number of analytical uses of this Hicksian construction remains amazing. Generations of students have learned their macroeconomics by mastering the standard IS–LM exercises: a decline in the marginal efficiency of capital shifts IS leftwards and thus reduces both income and the interest rate; an increase in the money supply shifts LM rightwards and hence raises income while lowering the interest rate, etc. But while Hicks succeeded in compressing a lot of macroeconomics into these two dimensions, such simplification naturally came at a cost. In the later literature, it has often proven difficult to keep the inevitable limitations of the apparatus in clear perspective. Three sets of such problems deserve mention.

The stock-flow dimensional aspects of the model is one problem area and the one with which Hicks himself has been most preoccupied. The IS-schedule is a locus of alternative *flow*-equilibria. But a flow-equilibrium has to be defined over some interval of time and in the case of production this interval has to be fairly long – Hicks (1983) suggests that we think of it as a 'year'. The LM-schedule shows alternative *stock*-equilibria. These can be defined for a point in time. But to insist that realizations stay consistent with expectations so that stock-equilibrium is maintained over an entire 'year', leaves so little uncertainty in the model that the demand for liquidity part of the LM-curve becomes difficult to rationalize. Hicks, therefore, sees a basic tension in the IS–LM construction between the periods appropriate to the two reduced forms. In still another reappraisal of IS–LM, Hicks (1986) points out that the IS–schedule summarizes the behaviour of the industrial sector and the LM that of the financial sector of the economy. This means that one of the characteristic simplifications of IS–LM analysis is, in effect, to disregard the balance-sheet of the industrial sector and the income-statement of the financial sector. This tends to direct attention away from cases where the two schedules are interdependent as, for instance, when an increase in current production is financed by bank credit so as to increase the money stock. It also makes IS–LM rather unsuitable for the analysis of many balance of payments problems.

A second set of problems has to do with a curious tendency for reliance on IS–LM to end up in a confusion of *nominal* and *real* (particularly real *intertemporal*) maladjustments. Consider, for instance, how the Phillips curve was used at one time to determine the expected pricelevel/ output composition of a change in the level of money income. But a change in money income of some given magnitude can be brought about by either a real disturbance (an IS-shift) or a nominal disturbance (an LM-shift). Why the price/quantity 'trade-off' of a nominal shock should ever have been expected to be the same as for a real shock appears in retrospect as a riddle but the nominal/ real distinction was seldom appropriately drawn in the Phillips curve controversy.

The most important instance of this nominal/ real confusion, however, occurred much earlier and is embedded in the outcome of the 'Keynes and the Classics' debate. This debate made use of IS–LM by investigating what restrictions on the static form of the model would produce an unemployment solution. Initially, two hypotheses seemed of interest (e.g. Modigliani 1944). One had the exogenous money supply set too low for nominal income to reach the level required for full employment at the given rigid money wage. the other had the liquidity preference function so specified that it kept the interest rate above the level required for saving and investment to be coordinated at full employment. In the course of the debate, however, this second possibility was

eliminated. It was pointed out by Pigou, Patinkin and others that, if it were possible to reduce the money wage rate to an arbitrarily low value, then the so-called 'Pigou effect' would reduce the propensity to save to whatever extent is necessary in order to bring saving and investment into line at full employment no matter what the level of interest rate happens to be. This Pigou-effect argument was taken to dispose of Keynes's intertemporal disequilibrium case. Thus the debate came to the distinctly odd conclusion that Keynes had revolutionized economic theory by asserting the classic platitude that when money wages are too high for equilibrium in the labour market unemployment is the result.

Losing sight of Keynes's intertemporal coordination failure ('saving exceeds investment') has proved costly to the Keynesian tradition. It deprived Keynesians of their natural response to Friedman's hypothesis that the lag in the adjustment of money wages is the only obstacle to employment finding its natural rate. Instead of pointing out that the natural rate of unemployment hypothesis is true only when the interest rate equates saving and investment at full employment, Keynesians tended to reply that money wages are even less flexible than Monetarists would like to believe. But if unemployment is due to money wages being too high in relation to the money supply and if one cannot afford to wait for wages to adjust, inflation will come to seem the normal remedy for unemployment. This tendency is reinforced by the fact that the rationale for the traditional Keynesian fiscal remedies is also lost when removed from the original context of intertemporal coordination failure. If saving exceeds investment (so that there is an excess supply of present resources and an implicit excess demand for future ones) it makes sense for the government to spend now and tax later. If, however, it can be presumed that real interest rates efficiently coordinate intertemporal activities, it may also be presumed that Ricardian equivalence is likely to hold.

A third problem area concerns the precise nature of the 'short-run' for which the IS–LM 'equilibrium' is defined. Comparative statics exercises with IS–LM often produce solution states that obviously represent situations of *incomplete* adjustment. The assumptions that might motivate such incomplete adjustment are, however, often unstated and not always obvious. By varying the information assumptions of the model, it is easy, for instance, to make both schedules shift in response to the standard disturbances. Such interdependence rather undermines the basic modelling strategy which presumes that one reduced form will stay put while the other shifts, so that the IS–LM diagram can be used to generate predictions in the same straightforward way as the Marshallian supply and demand apparatus. Moreover, when the information assumptions are not spelled out, confusion easily arises over whether elasticities or interdependence of the schedules are at issue in a particular controversy (Leijonhufvud 1983).

These points may be illustrated with reference to the theory of the monetary transmission mechanism. In vintage Keynesian theory, an increase in the money supply would shift LM rightwards while IS stayed put. If the interest elasticity of LM was high and that of IS low, monetary policy was seen to be 'ineffective' in the sense that, 'in the short run', the change in money income would be small. In a rational expectations model, a fully anticipated monetary impulse shifts both IS and LM in parallel fashion. In the short run, money income increases in full, constant-velocity proportion to the money injected into the system and this takes place independently of the interest elasticities of the two schedules. The older, Keynesian version can be rationalized in two distinct ways. Either one maintains that agents do not know of the nominal impulse or otherwise are unable fully to anticipate its eventual consequences; or else one interprets the increase in money, not as a pure nominal shock, but as an expansion of bank credit and inside money in, for instance, a regime with convertible money. IS–LM by itself, of course, will not help to settle the matter.

See Also

- ▶ Neoclassical Synthesis
- ▶ New Classical Macroeconomics

Bibliography

Gordon, R.J. (ed.). 1973. *Milton Friedman's monetary framework: A debate with his critics*. Chicago: University of Chicago Press.

Hansen, A.H. 1953. *A guide to Keynes*. New York: McGraw Hill.

Hicks, J. 1937. Mr Keynes and the classics: A suggested interpretation. *Econometrica* 5(April): 147–159.

Hicks, J. 1983. IS–LM: An explanation. In *Modern macroeconomics*, ed. J.P. Fitoussi. Oxford: Blackwell.

Hicks, J. 1986. Towards a more *general theory*. Paper delivered at a monetary theory symposium. Taipei, Taiwan, January.

Klein, L.R. 1947. *The Keynesian revolution*. New York: Macmillan.

Leijonhufvud, A. 1983. What was the matter with IS–LM? In *Modern macroeconomics*, ed. J.-P. Fitoussi. Oxford: Blackwell.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12: 45–88.

Modigliani, F. 1963. The monetary mechanism and its interaction with real phenomena. *Review of Economics and Statistics* 45(1), pt. 2: 79–107.

Patinkin, D. 1948. Price flexibility and full employment. *American Economic Review* 38(September): 543–564.

IS–LM in Modern Macro

Edward Nelson

Abstract

The IS–LM framework is associated with traditional macroeconomics, but versions of IS and LM functions can be justified using dynamic general equilibrium models that assume optimizing behaviour on the part of the private sector. The baseline version of these optimizing IS–LM relationships is discussed. Relative to the traditional IS–LM specification, the IS relationship in the optimizing IS–LM framework involves an extra term, which reflects the dependence of real aggregate demand on the expected level of spending next period. This extra term is implied by the intertemporal behaviour of households.

Keywords

Aggregate demand; Dynamic stochastic general equilibrium (DSGE) models; Infinite horizons; IS–LM in modern macro; IS–LM model; Monetarism

JEL Classifications

E1

Background: Traditional IS–LM

Some discussions use the term 'IS–LM' as a catch-all label for the approach of traditional Keynesian economics. The treatment here, however, will follow Sargent (1987, p. 53) in interpreting 'IS–LM' narrowly (and literally) as a pair of structural equations describing real aggregate spending as a function of the real interest rate, and real money demand as a function of scale and opportunity cost variables. From that perspective, it is not strictly accurate to talk of an 'IS–LM model' (since IS–LM is only a portion of a macroeconomic model) or to refer to the 'sticky-price assumption' of IS–LM (properly specified IS and LM equations are structural and should be independent of what is assumed about price behaviour, whose specification belongs to the supply side of a model; and IS–LM analysis in conjunction with price flexibility was considered even in Hicks 1937). Nor should models that have separate equations describing consumption and investment behaviour be considered models containing IS–LM equations, since deriving an IS equation necessarily involves eliminating the components of aggregate demand in favour of an expression for total demand.

It follows that, to find descendants of traditional IS–LM in modern macroeconomics, one should focus on cases where general equilibrium models produce a pair of equations clearly recognizable as corresponding to IS (real aggregate spending) and LM (real money demand)-type relationships.

IS–LM in Modern Macro: Early Literature

Early attempts to link IS–LM with dynamic optimizing macroeconomics include Aiyagari and Gertler (1985) and Fane (1985). These attempts, however, did not use infinite-horizon agents (the standard assumption in modern macroeconomics) and usually left more endogenous variables than output and the real interest rate in the equation for total spending, so this equation was not clearly recognizable as an IS relationship.

This early literature did show that it was possible to derive a conventional money demand equation from an optimizing model. This was also shown by McCallum and Goodfriend (McCallum and Goodfriend 1987) using an infinite-horizon model. A semilogarithmic version of McCallum and Goodfriend's money demand equation is:

$$rm_t = c_1 c_t + c_2 R_t \tag{1}$$

where $c_1 > 0$, $c_2 > 0$, and rm_t and c_t denote log-deviations of real money balances and real household consumption from their respective steady-state levels, with R_t being the short-term net nominal interest rate minus its steady-state value.

In light of the feasibility of deriving an LM relationship from an optimizing general equilibrium model, most discussions concentrated on whether IS-type relationships, and therefore IS–LM as a whole, are compatible with optimizing behaviour. A symposium on the subject of IS–LM and modern macroeconomics (Young and Zilberfarb 2000), which largely predated the recent literature, was generally negative about the prospects of linking up modern macroeconomics with IS–LM.

IS–LM in Modern Macro: Later Literature

In discussing the recent literature, it is worthwhile first stepping back to Hall (1978, p. 974), who showed that an infinite-horizon dynamic general equilibrium model implied an equation for aggregate household consumption (C) of the form

$$C_t^{-(1/\sigma)} = \beta(1 + r_t)\mathrm{E}_t C_{t+1}^{-(1/\sigma)} \tag{2}$$

where r_t is the short-term net real interest rate, β is the household's discount factor and $\sigma > 0$ is a utility function parameter (with a large σ value implying high intertemporal substitution in consumption). A log-linearized version of this equation is:

$$C_t = -\sigma(r_t - \mathrm{E}(r)) + \mathrm{E}_t c_{t+1} \tag{3}$$

where $\mathrm{E}(r)$ denotes the steady-state value of r_t. Consumption equations such as (3) continue to be present in the dynamic stochastic general equilibrium models prevalent today. What is different in the recent literature is a change in emphasis in interpreting the equation. Hall (1978) treats the real interest rate as fixed and focuses on the implied univariate behaviour of consumption. The recent literature does not treat the real interest rate as fixed, and instead builds up from the consumption condition (3) to an economy-wide description of aggregate real spending behaviour. If consumption is the only component of aggregate demand (implying the relation $c_t = y_t$), then eq. (3) implies an aggregate relationship – the optimizing IS equation – of the form:

$$y_t = b_1(R_t - \mathrm{E}_t \pi_{t+1}) + \mathrm{E}_t y_{t+1} \tag{4}$$

where $b_1 = -\sigma < 0$, π_t is inflation minus its steady-state value, and the Fisher condition $(r_t - \mathrm{E}(r)) = R_t - \mathrm{E}_t \pi_{t+1}$, has also been substituted in. Under the same approximation that output equals consumption, log output becomes the scale variable in the money demand function, so that eq. (1) implies an LM relationship:

$$rm_t = c_1 y_t + c_2 R_t. \tag{5}$$

Alternative assumptions to that of strict equality between consumption and output will deliver much the same IS relationship as eq. (4). For example, one could assume constant but non-zero investment, or random-walk exogenous investment behaviour (as in McCallum and Nelson 1999), or proportionality between

consumption and investment, and in each case derive an IS equation isomorphic to eq. (4).

Whatever the precise derivation, the common element in the recent literature that starts from the Euler consumption condition is that, instead of making restrictions, as Hall did, that lead to conclusions about the unforecastability of consumption growth, it sees the condition as underpinning a structural relationship describing the level of total real spending. In this relationship total spending is a function of the real interest rate, expected future spending levels, and exogenous shocks. The negative coefficient on the real interest rate allows a parallel with the traditional interest-elastic IS relationship $Y = f(r)$. That parallel has been highlighted by the later IS–LM literature, including Koenig (1989, 1993), McCallum (1989, p. 105), Woodford (1995, 2003), Kerr and King (1996), Rotemberg and Woodford (1997), and McCallum and Nelson (1999).

Shocks

It is straightforward to justify the addition of exogenous shock terms to the optimizing IS and LM equations. Preference shocks in the household's utility function can deliver this result: for the IS equation the shock is to the marginal utility of consumption; the LM shock, on the other hand, is a combination of the shocks to the marginal utility of consumption and to the marginal utility of services generated by real money balances. In addition, the portion of government spending that is not well approximated by a (log) random walk will produce a further rationale for an IS shock.

Treatment of Capital

As noted above, a restrictive assumption about investment (that is, that it is constant or random-walk exogenous) is needed to derive the optimizing IS eq. (4). Dupor (2001) criticizes such approximations on the grounds that investment is a sizable portion of aggregate demand and a major contributor, in arithmetic decompositions, to real GDP fluctuations. These facts can be accommodated, however, without making investment endogenous. One can simply assume that investment has a random-walk and a stationary component, both exogenous. The exogenous stationary component becomes a further IS shock and can be assumed to be highly variable.

General Equilibrium Status

Early discussions of dynamic general equilibrium models stressed the interdependence of aggregate demand and supply relationships, and, that being so, the infeasibility of labelling a subset of equations specifically 'aggregate demand' equations (see, for example, Sargent 1982). By contrast, the approach that derives IS and LM relationships from a general equilibrium analysis emphasizes that a subset of equations may be labelled 'aggregate demand' relationships; and that other conditions describing private sector behaviour (such as firms' pricing and hiring decisions and households' labour supply condition) constitute the aggregate supply block. Common separability assumptions regarding the private sector's preference and cost functions justify this division of equations. The central assumption is that the terms involving consumption, leisure, and real balances are additively separable in the private households' utility functions.

Is it an IS Equation?

The optimizing IS function has been criticized as not descriptive of the original 'investment–saving' acronym, as its baseline version comes from a model with no investment fluctuation and with saving zero or constant in equilibrium. But 'IS' was always a description of how aggregate spending fluctuations related to interest-rate variations – an 'income sensitivity' or 'interest sensitivity' equation rather than really an 'investment–saving' relationship. Detailed discussion of saving issues typically would not use the assumptions (for example, those regarding

infinite horizons for agents) underpinning baseline optimizing macroeconomic models.

Alternatively, the old 'investment–saving' label could be justified on the grounds that the IS equation forms part of a model describing the process by which investment and saving are equated. That description remains true of the optimizing IS equation; it happens that in the baseline model underlying this equation, the equilibrium occurs with saving and investment at constant or zero values.

Other Interest Rates

It is tempting to suggest that the optimizing IS eq. (4) is subject to the monetarist critique of traditional IS–LM because it excludes money from the IS equation. But in fact monetarists did not argue that money belonged in the structural IS equation. Instead, they argued that many yields mattered for aggregate demand and that these yields could not be summarized by a single interest rate (see, for example, Brunner and Meltzer 1973). Variations in money acquired significance because this spectrum of yields also appeared in the money demand function. The monetarist critique amounts to the suggestion, first, that different financial assets are not perfect substitutes, and second, that the discrepancies between the yields might be related to the behaviour of money. Baseline IS–LM, both old and new, presumes perfect substitutability between assets, in which case the short-term real interest rate is tightly related to other real returns prevailing in the economy. McCallum and Nelson (1999) defend the perfect-substitution assumption as the appropriate benchmark for many purposes. Nevertheless, as Bernanke and Reinhart (2004) argue, for some policy issues this assumption is not appropriate and so it would be desirable to break the link between different returns on assets, and investigate the effect of monetary policy actions on various yields. Such a generalization of IS–LM would tend to put extra real yields into the IS equation and extra nominal yields into the LM function.

See Also

- ▶ Hicks, John Richard (1904–1989)
- ▶ IS–LM

Bibliography

Aiyagari, S.R., and M. Gertler. 1985. The backing of government bonds and monetarism. *Journal of Monetary Economics* 16: 19–44.

Bernanke, B.S., and V. Reinhart. 2004. Conducting monetary policy at very low short-term interest rates. *American Economic Review* 94: 85–90.

Brunner, K., and A.H. Meltzer. 1973. Mr. Hicks and the 'monetarists'. *Economica* 40: 44–59.

Dupor, B. 2001. Investment and interest rate policy. *Journal of Economic Theory* 98: 85–113.

Fane, G. 1985. A derivation of the IS–LM model from explicit optimizing behavior. *Journal of Macroeconomics* 7: 493–508.

Hall, R.E. 1978. Stochastic implications of the life-cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy* 86: 971–987.

Hicks, J.R. 1937. Mr. Keynes and the 'Classics': A suggested interpretation. *Econometrica* 5: 147–159.

Kerr, W., and R.G. King. 1996. Limits on interest rate rules in the IS Model. *Federal Reserve Bank of Richmond Economic Quarterly* 82: 47–75.

Koenig, E.F. 1989. A simple optimizing alternative to traditional IS–LM analysis. Manuscript, *Federal Reserve Bank of Dallas*.

Koenig, E.F. 1993. Rethinking the IS in IS–LM: Adapting keynesian tools to non-Keynesian economies, part 1. *Federal Reserve Bank of Dallas Economic Review* 78: 33–49.

McCallum, B.T. 1989. *Monetary economics*. New York: Macmillan.

McCallum, B.T., and M.S. Goodfriend. 1987. Demand for money: Theoretical studies. In *The new palgrave: A dictionary of economics*, vol. 1, ed. J. Eatwel, P. Newman, and M. Milgate. London: Macmillan.

McCallum, B.T., and E. Nelson. 1999. An optimizing IS–LM specification for monetary policy and business cycle analysis. *Journal of Money, Credit and Banking* 31: 296–316.

Rotemberg, J.J., and M. Woodford. 1997. An optimization-based econometric model for the evaluation of monetary policy. *NBER Macroeconomics Annual* 12: 297–346.

Sargent, T.J. 1982. Beyond demand and supply curves in macroeconomics. *American Economic Review* 72: 382–389.

Views expressed in this paper are the author's and should not be interpreted as those of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

Sargent, T.J. 1987. *Macroeconomic theory*, 2nd ed. New York: Academic.
Woodford, M. 1995. Price-level determinacy without control of a monetary aggregate. *Carnegie-Rochester Conference Series on Public Policy* 43: 1–46.
Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.
Young, W., and B.Z. Zilberfarb (eds.). 2000. *IS–LM and modern macroeconomics*. Boston: Kluwer.

Isnard, Achylle Nicolas (1749–1803)

R. F. Hébert

Keywords

Boisguilbert, Pierre le Pesant, Sieur de; Capital theory; General equilibrium; Isnard, A.N.; Mathematical economics; Optimum resource allocation; Physiocracy; Simultaneous equations; Value; Vauban, S. le P.; Walras, L.

JEL Classifications

B31

French engineer and economist, Isnard was born at Paris on 25 February 1749; he died at Lyons on 25 February 1803. There are no details of his family history except that he had a devoted brother, J.L. Isnard, who was a lawyer and a judge, and who often interceded on his behalf. At the age of 17, Isnard entered the Ecole des Ponts et Chaussées which, even at this early date, inspired interest in political economy and exposed its students to heavy doses of mathematics and statistics. On successfully completing his studies, Isnard began his career as an apprentice engineer in the district of Besançon. While engaged in various works of construction in these environs, he took the time to write his remarkable two-volume work, *Traité des richesses,* which was published in 1781.

Isnard's *Traité* is a highly original work, despite the fact that its theoretic core is embedded in otherwise unexceptional arguments against Physiocratic doctrines. By this fact, we may infer that Isnard knew the Physiocratic literature, but we can only speculate on his acquaintance with other writers. Given his background and training, the authors he would have most likely known are Boisguilbert and Vauban (Boisguilbert's ideas were represented in 19th-century course outlines at the Ecole des Ponts et Chaussées, and Vauban's views on the professionalization of engineers were largely responsible for the establishment of the Ecole). Boisguilbert certainly had a vision of an interconnected economy and of a kind of general equilibrium, although he failed to render his conception concrete by erecting any kind of formal, theoretic structure of a mathematical nature.

Isnard, on the other hand, was the first writer to attempt a mathematical definition and a mathematical proof of an economic equilibrium. Furthermore, he gave specific form to the general equilibrium concept by constructing a set of simultaneous equations which, in general form and content, anticipated the major elements of the Walrasian system, including the general interdependence of markets and quantities, the technical specifications of the exchange ratios, and the mathematical determination of the *numéraire.* It remained for Walras to add the engine of utility maximization and to adapt Isnard's model to his own purposes, something which, according to Jaffé (1969), he did with persistence, if not with ease. Isnard's pioneer efforts do not in any way denigrate Walras's monumental achievement, but they do lend force to the conviction that the development of economics was, and remains, a cumulative process.

Isnard's *Traité* is now extremely rare. However, the mathematics of his equilibrium analysis of exchange are partially accessible in Robertson (1949), Baumol and Goldfeld (1968), Jaffé (1969), and Theocharis (1983). The significance of Isnard's performance is that he discovered early on the truth that value is not an intrinsic thing but rather is a magnitude which necessarily varies in relation to other goods, whose worth is also interdependent. Specifically, Isnard anticipated the two- good world of Walras in which, for example, the demand for eggs is the supply of

wheat and the demand for wheat is the supply of eggs. This elaboration of commodity interdependencies in real terms consumes approximately the first half of Walras's *Eléments*. Mathematically, Isnard treated value as an exchange ratio, moreover, and he worked out the equilibrium process of exchange both with and without money.

It is noteworthy that Isnard extended the subjects under his analytical purview to include, besides the theory of exchange, the theories of production, capital, interest and foreign exchange. Jaffé (1969) has demonstrated that Walras's economic theory bears the imprint of Isnard in each of these areas. Underscoring merely the most striking example of the calibre of Isnard's analysis, Jaffé (1969, p. 40) emphasized his theory of capital and interest, which correctly laid down the rule for optimum resource allocation in the following terms:

> Capitals are distributed among different employments in agriculture, industry, and commerce in such a way that the ratios of their values to receipts from the sale of their products less the costs of upkeep, repair, and replacement – that is, the ratios of [invested] funds to [net] returns – are everywhere the same in all enterprises. This uniformity is achieved and equilibrium established because funds flow to and abound in places where the yield [intérêt] is highest and because like things have one and the same value. When things have a higher price in one place than in another, they rush there and equilibrium is re-established. Let F be the value of the funds employed in agriculture and F′ that of the funds employed in industry; let B be the payments for the value of the products of agriculture less the cost of upkeep, repairs and replacement and B′ the payments for the products of industry less the same costs, then the ratio of F to F′ must be equal to the ratio of B to B′ for the ratio of F to B to be equal to the ratio of F′ to B′ or for the rate of interest [in the sense of rate of capitalization] to be everywhere the same. This uniformity [in the rate of capitalization] is realized not only between agriculture and industry in general, but also among individual enterprises.

It is, of course, necessary that perfect knowledge obtain for this conclusion to hold, but even without always making his assumptions explicit, Isnard anticipated much of modern microeconomic theory.

The scope and sweep of his analysis unquestionably entitle Isnard to a position of prominence in the history of economic thought. Yet appropriate recognition took a long time. Despite the filiation of ideas between Isnard and Walras, the 'father' of general equilibrium analysis mentioned Isnard's name in only one place, and that an obscure bibliographic article (a French reprint of Jevons's famous bibliography of mathematico-economic works) published in the *Journal des Economistes* in 1878. Add to this the ambiguity, idiosyncrasy and prolixity of Isnard's treatise. His definition and use of mathematical symbols is inconsistent and the essence of his arguments difficult to extract, nested as they are in a morass of other material that is neither very original nor very interesting. Such deficiencies were bound to handicap the recognition and acceptance of Isnard's contribution. In the final analysis, however, Isnard was simply a brilliant pioneer who wrote ahead of his time, and like so many other semi-tragic heroes of economic analysis (for example, Cournot and Gossen), he failed to receive his due until long after departing the scene.

Isnard suffered in his personal life even as his ideas suffered (by neglect) in economics. Hot-tempered, yet not given to the intrigues apparently required to advance in the engineering ranks of a quasi-military public service, Isnard spent most of his career in a subordinate capacity. After he finally received a post worthy of his talents, his wife died, leaving him to raise three motherless children. At that point Isnard left government service and struggled in penury for some time. Recalled by Napoleon for the Egyptian campaign in 1798, he was inexplicably left behind. Adding insult to injury, he was forced to take an oath of allegiance to the Republic even though he was an avowed royalist. He later became a member of the Tribunate under Napoleon and took an active part in the formation of public finance and conscription policies. But upon completing his term he resumed his engineer's career at Lyons, where he died soon after, 54 years to the day from his birth. Given his apparent influence on Walras, Theocharis (1983, p. 62) probably did not exaggerate much when he labelled Isnard's *Traité* 'one of the most important contributions in the history of the development of mathematical economics'.

I

Selected Works

1781. *Traité des richesses,* 2 vols. London/ Lausanne: F. Grasset.
1801. *Considérations théoriques sur les caisses d'amortissement de dette publique*. Paris: Duprat.

Bibliography

Baumol, W.J., and S.M. Goldfeld, eds. 1968. *Precursors in mathematical economics: An anthology.* London: London School of Economics and Political Science.
Jaffé, W. 1969. A.N. Isnard, progenitor of the Walrasian general equilibrium model. *History of Political Economy* 1: 19–43.
Robertson, R.M. 1949. Mathematical economics before Cournot. *Journal of Political Economy* 57: 523–536.
Theocharis, R.D. 1983. *Early developments in mathematical economics*. 2nd ed. London: Macmillan.

Italy, Economics in

Massimo M. Augello and Marco E. L. Guidi

Abstract

The history of economics in Italy reflects an interaction between scientific–educational institutions and political power, which led economists to combine a theoretical approach and political commitment. During the Enlightenment a network of circles and public academies spawned the contributions of Beccaria, Genovesi, Galiani, Ortes and Verri. The institutionalization of economics in the 19th century prepared the success of the marginalist generation led by Pantaleoni, Pareto and Barone. In the interwar period, academic economists formed a bulwark against Fascism. The post-war political climate favoured the internationalization of economics, with the importation of Keynesianism and, later, other currents of thought.

Keywords

Barone, E.; Beccaria, C.; Boccardo, G.; Bodio, L.; Boselli, P.; Botero, G.; Bresciani Turroni, C.; Caffè, F.; Carey, H. C.; Cognetti de Martiis, S.; Corbino, E.; Corporatism; Cossa, L.; Cost of reproduction theory of value; Croce, B.; Custodi, P.; Davanzati, B.; De Viti De Marco, A.; Del Vecchio, G.; Demaria, G.; di Fenizio, F.; Division of labour; Einaudi, L.; Taviani, P. E.; Engel, E.; Entrepreneurship; Equilibrium; Fanfani, A.; Fanno, M.; Ferrara, F.; Fuà, G.; Galiani, F.; Galilean revolution; Garegnani, P.; Genovesi, A.; German Historical School; Gioja, M.; Graziadei, A.; Graziani, A.; Intieri, B.; Italy, economics in; Jannaccone, P.; Keynesianism; Labour theory of value; Labriola, A.; List, F.; Loria, A.; Luzzatti, L.; Machiavelli, N.; Marginalism; Marrama, V.; Mazzola, U.; Messedaglia, A.; *Methodenstreit*; Modigliani, F.; Montanari, G.; Morpurgo, E.; Nazzani, E.; Neoclassical synthesis; Neo-Ricardian economics; Nitti, F. S.; Ortes, G.; Pantaleoni, M.; Pareto, V.; Pasinetti, L.; Pesenti, A.; Physiocracy; Platonism; Protonotari, F.; Public finance; Quetelet, A.; Rae, J.; Revisionism; Ricci, U.; Romagnosi, G.D.; Saraceno, P.; Say, J.-B.; Scaruffi, G.; Scialoja, A.; Serra, A.; Sharecropping; Sismondi, J. C. L. S. de; Smith, A.; Società Italiana degli Economisti; Sraffa, P.; Steve, S.; Subjectivist theory of value; Supino, C.; Sylos Labini, P.; Toniolo, G.; Turati, F.; Vanoni, E.; Vasco, G.; Verri, P

JEL Classifications

B1

This article examines the evolution of Italian economic thought from its origins to the post-Second World War years, when the professionalization and internationalization of economics reached maturity, sowing the seeds of the present vigorous state of economic studies. It offers an institutional history of political economy, distinguishing four epochs. The first epoch runs from the 16th to the 18th century. In this period, the alliances between

enlightened sovereigns and groups of intellectuals produced a wealth of original contributions to economics. The second epoch corresponds to the Napoleonic age and the Restoration. This was a period in which, despite political repression in Italy, Smithian political economy penetrated many circles and was debated in journals and academies. The third epoch runs from the unification of Italy in 1860 to the rise of the Fascist regime in 1922, and is referred to as the 'liberal age'. This period was crucial for the institutionalization of economics and for the prominent public role that was attributed to economists. Such a favourable environment was responsible for the high standard of scientific debate, which culminated in the generation of Pantaleoni and Pareto, when Italian economics, as Schumpeter wrote (1954, p. 855), 'was second to none'. Finally, the fourth epoch regroups the Fascist era and the post-war years. During the decades of Fascism, the regime's attempts to control economic debate produced a reaction of self-defence and isolation among neoclassical economists. After the war, the new climate of liberty encouraged economists to get back to their public role. The political debate on economic planning was responsible for the acceptance of Keynesianism in the 1950s. The success of Neo-Ricardianism in the 1960s and, later, of American-based mainstream economics marked the internationalization of Italian economics.

Public Happiness and Geometrical Method: From the Origins to the Enlightenment

Although in the Middle Ages theological debate over the 'just price' and the legitimacy of usury flourished in many parts of Italy, and Italian authors came to be respected throughout Europe, the beginnings of modern economic science in this country date from the 16th and early 17th centuries, when the formation of regional states generated a need to regulate public finances, trade and the circulation of money. Some traces of economic analysis can be found in the treatises on politics published by the humanist Niccolò Machiavelli (1469–1527) *(Discorsi sopra la prima deca di Tito Livio*, 1513–21) and the Jesuit Giovanni Botero (1544–1617) *(Della Ragion di Stato,* 1589). But the most original analyses were contained in some short treatises of a systematic and highly formalized character, based on a hedonistic framework, a coherent theory of value and an extensive use of mathematics. The theoretical framework they provided could be employed to interpret the monetary and commercial problems of the time. Of this kind are the *Discorso sulle monete* (1582) by the aristocrat Gaspare Scaruffi, the *Lezione delle monete* (1588) by the merchant and historian Bernardo Davanzati (1529–1606), the *Breve trattato delle cause che possono far abbondare li regni d'oro e d'argento, dove non sono miniere* (1613) by Antonio Serra, whose life is shrouded in mystery, and the *Trattato mercantile della moneta* (1683) by Geminiano Montanari (1633–87), a professor of mathematics and astronomy. These works were connected to the diffusion of the Catholic currents inspired by Platonism and hostile to Aristotelian scholasticism that were at the root of the Galilean revolution. The abstract nature of these texts can be explained by the fact that they were a product of the scientific academies created in this period under the aegis of the Italian princes – especially those in Florence – with a view to encouraging knowledge which could be useful in strengthening state power and countering the civil decay of the country, for which they blamed the political influence of the Church. The authors of these economic treatises were natural philosophers who acted as temporary consultants to government or were members of the state bureaucracy.

The long wave of Galilean and Platonic doctrines continued through the 18th century and had a recognizable impact on the theoretical structure of the economic discourse in the two circles that were at the centre of the Italian Enlightenment, combining in different ways with new ideas coming from France and Scotland and with ideas from other indigenous traditions. The first of these circles was the Accademia dei Pugni of Milan, which was run by Pietro Verri (1728–97), author of the *Meditazioni sull' economia política* (1771) and Cesare Beccaria (1738–94), known not only for his main work, *Dei*

delitti e delle pene (1764), but also for a series of theoretical articles on political economy published in the journal of the academy, *Il Caffè* (1764–6). The second group was that in Naples headed by Bartolomeo Intieri, whose discussions sparked both *Della moneta* (1751) by Ferdinando Galiani (1728–87) and *Lezioni di economia civile* (1766–7) by Antonio Genovesi (1713–69). These groups became involved in the economic reforms their monarchs were trying to bring about, as revealed by the public offices obtained by Verri, Beccaria and others. Also in other parts of the country, the academies were called upon to produce a science whose utility was measured in terms of greater dominion over nature and of the greater ability of governments to increase 'public happiness'. But it was the 1753 foundation of the Accademia Economico-Agraria dei Georgofili in Florence that sanctioned the scientific status of economics and its political role in the strategy of reform. Right up to the end of the 18th century, this academy was to be one of the main vehicles for the spread of Physiocratic and Smithian doctrines in Italy. Its journal was an example of the many agricultural periodicals that hosted economic debates, often of a practical kind, albeit open to the new science of political economy.

Collaboration between philosophers and princes ushered in the creation of the first teaching of political economy as part of a reform of university studies whose purpose was to bring them under the umbrella of the state, combating the control that religious orders and professional bodies had traditionally exerted over them. A chair of Commercio e meccanica was founded in Naples in 1754 on the initiative (and funding) of Intieri, and conferred on Genovesi. Another professorship of political economy, established at the Scuole Palatine, Milan, in 1768, was assigned to Cesare Beccaria. Similar chairs were created in Modena, Catania and Palermo. The aim of these chairs was twofold: they should instruct government bureaucrats in the art of governing economic and financial affairs, and stimulate the application of new agricultural techniques and agrarian laws.

From a theoretical point of view, the contribution of these authors – to whom one should add at least Giammaria Ortes (1713–90) and Gianbattista Vasco (1733–96) – is quite homogeneous. Their core approach is based on a natural law framework which turns around a static rather than dynamic notion of equilibrium, and suggests that there are forces in society that tend to restore equilibrium when natural disasters, changes in tastes, or political errors create unbalances. Another basic assumption is the sensationalist view that human beings are constantly under the guidance of pleasure and pain. The underlying methodology is still abstract and 'geometrical', as in the works of their predecessors. The focus of analysis is on problems of exchange rather than of production (although Beccaria gave the clearest definition of the division of labour before Adam Smith). Their main contributions concern the analysis of value, based on utility and scarcity: Ortes, Beccaria and Verri attempted a mathematical formulation of the law of demand and supply as a guide to the analysis of price adjustments. Another theme of inquiry is the theory of money, where Galiani, adopting metallist assumptions, expounded a clear distinction between short-run variations in the value of money and long-term effects, and between real and monetary effects. Finally, these authors shared a view of the sovereign as a reformer and supreme moral authority, who takes into account the feelings and needs of individuals and constructs a social order according to the dictates of reason and natural law. This order consists of an equilibrium of interests that generates 'public happiness' *(felicità pubblica).*

The Spread of Classical Economics in the Age of *Risorgimento* (1815–60)

Although the Napoleonic age is considered, in Europe as a whole, a period in which political economy was regarded with suspicion, in Italy the establishment of the Empire's satellite kingdoms favoured the discipline's development and its institutionalization. However, the content of teaching was radically modified: theoretical economics was reduced in order to make room for legal and statistical notions, which were considered more urgent for the training of public officials. Moreover, Napoleonic administrations

concentrated on the collection of detailed statistical information about the condition of their *départements*, in order not only to promote anti-feudal reforms but also to protect French interests. A new generation of government officials was assigned to this task: among them there was the most important economist of this period, Melchiorre Gioja (1767–1829), who published his main work, the *Nuovo prospetto delle scienze economiche,* between 1815 and 1817. Gioja examined Smith's and Say's theories with a critical eye, and his original analysis of cooperation, division of labour and machinery was acknowledged by Charles Babbage as an anticipation of his own theories. Regarding economic policy, Gioja was favourable towards state intervention in order to foster the development of agriculture and manufactures.

Another Napoleonic official was Pietro Custodi (1771–1842), who from 1803 to 1805 edited the 50 volumes of a collection titled *Scrittori classici italiani d'economia politica,* which reproduced most of the Italian texts on political economy from previous centuries. Custodi aimed at stimulating the patriotic spirit of his fellow citizens by encouraging them to improve their economic and statistical knowledge. This collection produced in the next generation of intellectuals of the *Risorgimento* era an exaggerated feeling of national pride, which nevertheless encouraged the study of economics.

Economics experienced its worst period after the Restoration in 1815. The reactionary governments of the Italian regional states considered the teaching of economics to be a vehicle for liberal and democratic ideas. As a consequence, all chairs of political economy were suppressed, except in Naples and Sicily, where they were put under strict political control. Only in the 1840s, in Piedmont and Tuscany, with the establishment of constitutional governments, was the teaching of political economy restored. Antonio Scialoja (1817–77), first, and then Francesco Ferrara (1810–1900) were appointed professors at the University of Turin, while other chairs of economics were created in Pisa and Siena.

In these conditions, discourse on political economy went on largely outside universities. This does not mean that it was clandestine, since it was developed in academies and associations which enjoyed an official status. But the political control over these institutions implied that public debate on controversial issues was sometimes tolerated and sometimes heavily repressed. Already in the Napoleonic age newly founded institutions, such as the Accademia Pontaniana of Naples or the Istituto Nazionale, established in Bologna and transferred to Milan in 1810, had included departments of moral and political sciences where political economy was discussed. Furthermore, the experiences of 18th-century agrarian academies had prompted the establishment of a network of provincial associations termed 'agrarian' or 'economic societies', which aimed at promoting the development of local economies. These associations continued their activities even during the decades following the Restoration, expanding from Piedmont to Sicily. Despite their eminently practical goals, economic societies gave an important impetus towards the spread of British and French political economy and laissez-faire ideals.

Another means by which political economy spread through the Italian learned classes was the periodical press, despite the existence of censorship. In the early decades of the 19th century the heading 'political economy' appeared on a growing number of articles published in new journals of 'sciences, letters and arts', such as the *Biblioteca italiana,* founded in Milan in 1816, the *Antologia,* created in Florence in 1821, and *Il progresso delle scienze, delle lettere e delle arti,* first published in Naples in 1832, where it acted as the main point of convergence of liberal culture. Lively exchange of ideas was also found in journals of agriculture, especially the *Giornale agrario toscano,* founded in 1827, which together with the Accademia dei Georgofili promoted an original debate on sharecropping echoing Sismondi's remarks in *Tableau de l'agriculture toscane.*

The first signs of a trend towards specialization in economic disciplines came with the birth of several journals mainly devoted to statistical and economic themes, such as the *Annali universali di statistica* and the *Giornale di statistica.* The former was first published in Milan in 1822

and had among its contributors Gioja and Giandomenico Romagnosi (1761–1835). The latter was founded in Palermo in 1836 as the organ of the Central Statistical Office. Edited by Ferrara, it achieved immediate recognition as the premier forum for debate among Sicilian laissez-faire economists. Another interesting experiment was *Il Politecnico,* launched by Carlo Cattaneo (1801–69) in 1839. The majority of the essays were composed by Cattaneo himself, and dealt with various practical issues. However, in two remarkable articles Cattaneo focused on doctrinal questions, criticizing the protectionist theories of Friedrich List, and arguing that knowledge and motivation are the most important factors of economic development. But one should also mention a number of journals created in Naples in the 1840s, which arose against the backdrop of private law schools, established as an alternative to the more conservative form of instruction offered by the universities. These journals and institutions soon became a focal point for the new school of liberal economists, of which Scialoja was the main representative.

As this description makes clear, the political economy debated in these forums was that of Smith and Say. In northern Italy a key figure was Romagnosi, a legal philosopher who – taking inspiration from Giambattista Vico's philosophy of history – formulated a peculiar version of Smithian political economy, in which the notion of 'natural progress of opulence' was employed to argue that economic development depended on a framework of formal and informal institutions (so-called *incivilimento),* and that government-induced industrialization would result in social disaster. Romagnosi's 'institutionalist' approach influenced a whole generation of economists, stimulating interesting contributions on the relationships between law and economics. In the south of Italy, the penetration of classical economics was mediated by the influence of the French *idéologues,* which caused Say's work to be received enthusiastically. The most brilliant product of this environment was Scialoja's *I principj della economia sociale esposti in ordine ideologico* (1840), translated into French in 1844, which adopted Say's subjectivist approach to value and developed the analysis of the entrepreneur in a pre-Schumpeterian sense.

But the most acute and original economist of this age was Ferrara, who in his lecture notes of 1856–8 and in the prefaces to the *Biblioteca dell' economista* – a 'library' containing the Italian translation of a vast number of foreign works on economics, of which he was the editor – proposed a generalization of the cost of reproduction theory of value formulated by Henry Carey and John Rae. Ferrara's version of this theory took into consideration three different cases: that of 'physical reproduction' by direct labour, that of physical reproduction 'by way of exchange' and that of 'economic' reproduction by substitutes. In this way, the theory highlighted the fact that value is grounded on utility and subjective opportunity costs, clearly foreshadowing marginalist analysis.

The Institutionalization of Economics in the Liberal Age (1860–1922)

The epoch that followed the unification of the country in 1860 was decisive for the consolidation of economic studies. Chairs of political economy were introduced in the more than 20 law faculties that existed at the time. In 1876, new university regulations added the teaching of statistics and public finance. The latter was established as a compulsory course in 1885. Likewise in the 1880s, two Higher Schools of Commerce were created in Genoa and Bari, similar to the first institution of this kind, which had been founded in Venice in 1868. This expansion multiplied the opportunities for economists to obtain university positions, and well before the end of the 19th century the social identity of the economist could be largely identified with the academic profession. But a decisive stimulus to the professionalization of economics was provided in the mid-1870s by the explosion of the Italian counterpart of the *Methodenstreit*, the dispute over methods that divided German-speaking economics. All the major economists became involved in it, and opposition between different economic and political conceptions had an important impact on the professional and academic level. These

divisions induced economists to devote greater attention not only to the scientific aspects of the profession (training and specializations) but also to academic policy (increase in the number of academic chairs and control over recruitment procedures). This process resulted in a generational change within the ranks of academic staff, leading to a preponderance of the followers of *Kathedersozialismus* or 'socialism of the chair' in the German mould.

Another important element is represented by the increasing public role played by economists. The extension of civil liberties, coupled with the institution of a national representative system, gave them an extraordinary opportunity to spread economic knowledge and influence policymaking. Many economists became columnists for newspapers and weekly magazines, while others were active in the foundation of associations of interests, chambers of commerce, saving banks or cooperatives. Lastly, virtually all the leading economists of this age – more than 30 – became members of parliament. And although some of them were involved in parliamentary activities that bore little relation even to the broadest view of the scope of political economy, in the central debates on tariffs and trade, fiscal policies, credit, education, and in inquiries on the condition of agriculture and industry the voice of economists became a typical feature of public life. Some economists were also appointed ministers, while three of them – Paolo Boselli (1838–1932), Luigi Luzzatti (1841–1927), and Francesco Saverio Nitti (1868–1953) – became prime ministers.

On the whole, these activities strengthened the scientific and social identity of economists.

The growing professionalization of economists was also reflected in the creation of new societies in which they played a central role. After the first experiments in Turin in the 1850s and early 1860s under the guidance of Ferrara, the Società di Economia Politica Italiana was established in 1868 on the initiative of the economist Francesco Protonotari, who was the editor of the most important scientific and literary journal of the time, the *Nuova Antologia.* These associations attracted the great majority of academic economists and many representatives of the political elites. The constitution of the Società di Economia Politica significantly opened with the statement that the Society's mission was to 'promote and disseminate economic studies'. However, very soon its activities were dominated by more practical discussions on parliamentary debates and government economic policy. Conflicts concerning the new orientation of the society's purposes led to a gradual slowdown in the pace of activities.

It was against this backcloth that the *Methodenstreit* arose, breeding the projects of two rival associations. The former, dubbed Società Adamo Smith, was set up in Florence in 1874 on the initiative of Ferrara and several laissez-faire economists and politicians belonging to the group of the so-called 'Tuscan moderates'. The aim of the Society was that of 'promoting, developing and defending the doctrine of economic liberties', and of assuming the character of a scientific body, excluding from debate all that could be more properly described as political. The latter society, called Associazione per il Progresso degli Studi Economici, was created in January 1875 by a group headed by Luigi Cossa (1831–96), a powerful academic of the university of Pavia, Fedele Lampertico (1833–1906), an influential senator of the Venetian area, and by Luzzatti and Scialoja. Responding to Ferrara's splinter-group tendency, these economists had drawn up a document, known as the 'Padua circular', which marked the start of the counteroffensive by 'socialists of the chair'. The society set itself the task of promoting social studies, to be accomplished partly through extension of its organizational structure to different parts of Italy.

Both associations proved to be short-lived, but the overall effect of their activities over roughly a 30-year period was that of ushering in a profound change in the institutional set-up of economic studies, reinforcing the academic and public background of the economists' activities. This new condition was reflected in the world of publishing. To begin with, while scientific–literary journals continued in their tradition of hosting writings on economic themes, more specialized journals were established. Characteristically, in the 1870s almost all of the economic journals exhibited very close

links with one or the other of the conflicting schools of economic thought. Thus, orthodox liberals used as their mouthpiece the journal *L'Economista,* created in 1874, whereas 'socialists of the chair' founded one year later the *Giornale degli economisti*. Rather than a genuine forum for scientific debate, however, such journals tended to become tools with which to enter the political fray. This characteristic was to a lesser extent replicated by the new journals that appeared in the 1890s, despite their more scientific and academic nature: the most important among them were the socialist periodical *Critica sociale,* directed by Filippo Turati (1857–1932), the *Rivista internazionale di scienze sociali e discipline ausiliarie,* edited by the Catholic economist Giuseppe Toniolo (1845–1918), and *Riforma sociale* (1894–1935), edited at first by Nitti, and later by Luigi Einaudi (1874–1961). Perhaps the first journal of economics in the modern sense may be considered the new series of the *Giornale degli economisti,* started in 1890 and managed by Maffeo Pantaleoni (1857–1924), Antonio De Viti de Marco (1858–1943) and Ugo Mazzola (1863–99). This journal voiced the radical laissez-faire approach of its editors, while becoming the forum of academic research and the main vehicle of penetration of marginalist theory in Italy.

Even outside the world of journals, greater attention began to be paid to the promotion of economic studies. For instance, after the first two series of the *Biblioteca dell' economista,* edited by Ferrara and published in the 1850s and 1860s, a third series was entrusted to Gerolamo Boccardo (1829–1904) in the 1870s, and a fourth and fifth series were continued by Salvatore Cognetti de Martiis (1844–1901) and Pasquale Jannaccone (1872–1959) up to 1922. With its 71 volumes containing more than 150 classics of economics, the *Biblioteca* was extremely successful and became a unique tool for those who wanted to update their knowledge in this field. Economists also popularized their doctrines through dictionaries and encyclopaedias; the most important of them was the *Dizionario della economia politica e del commercio,* published in 1857–61 by Boccardo, who was also the editor of the *Nuova enciclopedia italiana* (1875–88).

Undoubtedly the main instrument for the spread and institutionalization of political economy was the large number of treatises, manuals and popularizations that were published during this period. The authors of these texts were both major economists and a host of lesser-known scholars, philanthropists and schoolteachers interested in the popularization of political economy. The number of works published – almost 300 from 1840 to 1920 – reveals that there was a pervasive 'need' for political economy, considered as a discipline that could educate the younger generations of administrators and politicians, instruct public opinion and enlighten the working classes. To judge from the number of editions, the most popular manuals were Cossa's *Primi elementi di economia politica* (1875, 17 re-editions), Emilio Nazzani's *Sunto di economia politica* (1873, 16), Boccardo's *Trattato teoricopratico di economia politica* (1853, nine), Camillo Supino's *Principii di economia politica* (1904, nine), Achille Loria's *Corso completo di economia politica* (1909, seven), and Augusto Graziani's *Istituzioni di economia politica* (1904, six).

At the same time, the scientific quality of the work was high. Italian economists rapidly assimilated international economic debates, and in some cases they became important protagonists. The quantitative approach to statistics initiated by A. Quetelet and E. Engel was largely accepted in the mid-1860s thanks to the contributions of Angelo Messedaglia (1820–1901), whose methodological works influenced a whole generation of economists, Emilio Morpurgo (1836–1885), and Luigi Bodio (1840–1920), who organized the Central Statistical Office and was elected secretary of the International Institute of Statistics on its foundation in 1885. Some years later, Cossa, Lampertico and Luzzatti were instrumental in familiarizing Italian scholars with the methodology of the German Historical School and the social views of socialism of the chair. These economists promoted a renewal of economic studies along inductivist and quantitative lines, and adopted a critical stance vis-à-vis economic liberalism in matters of social policy. They were called the 'Lombard–Venetian School' since most of them taught at the universities of Pavia and Padua.

An even more vigorous and original response to outside stimuli was represented by the penetration of marginalism. Pantaleoni's *Principii di economia pura* – a work largely inspired by Jevons, Edgeworth and Marshall – dated from 1889, but the same author had already published in 1883 a work on public finance based on marginalist notions. Pantaleoni encouraged Vilfredo Pareto's (1848–1923) conversion to the new approach some years later. In 1893, the latter succeeded Walras at the chair of economics in Lausanne. In his *Cours d'économie politique* (1896–7), and more radically in *Manuale di economia politica* (1906), he revolutionized utility theory, laying the foundations of modern microeconomic analysis. A third representative of Italian marginalism was Enrico Barone (1859–1924), whose article on 'The Ministry of Production in the Collectivist State' (1908) was included by Hayek in his 1935 anthology on economic planning. Interesting applications to public finance were also provided by Barone himself and by De Viti and Mazzola. Their contributions lay the foundation of an original school of thought whose analysis of taxes, public expenditure, and of the political context in which fiscal structures operate, has been recognized by James Buchanan as the starting point of the development of modern public finance theory. A distinctive feature of Italian marginalist economists was their practical and ideological commitment: they engaged themselves in political and editorial activities, staunchly defending a radical laissez-faire view. The socialists Arturo Labriola (1873–1959) and Enrico Leone (1875–1940) attempted to find a compromise between marginalism and Marxism. In the first decade of the 20th century, neoclassical economics had already become the orthodox approach.

Less vigorous, albeit no less original, was the Italian contribution to Marxist revisionism. In *La rendita fondiaria e la sua elisione naturale* (1880), Loria (1857–43) attempted to explain the functioning of a capitalist economy as a result of the structure and evolution of landed property. The historical and theoretical weaknesses of Loria's approach were then attacked at the end of the century by the Marxist philosopher Antonio Labriola (1843–1904), but his efforts to convince Benedetto Croce (1866–1952) to join his camp resulted in a relaunching of revisionism: Croce considered Marx's notion of surplus value as a simple 'mental abstraction' which could not explain the essence of capitalist production. On the other hand, Antonio Graziadei (1873–1953) argued that the labour theory of value was useless to explain the genesis of surplus value and the formation of market prices.

From Corporatism to Keynesianism and Neo-Ricardianism

After an early phase of authoritarian laissez-faire policy delegated by Mussolini to the economist and minister of finance Alberto De' Stefani (1879–1969), a turn towards a corporatist organization of the economy was accomplished in 1926. The introduction of corporatism was the result of political decisions rather than of scientific debate, although corporatist currents of Catholic and socialist ascendancy had existed since the late 19th century. The fascist regime organized two national conferences in 1930 and 1932 to stimulate a debate on corporatist economics, but they ended up with the defeat of those intellectuals who stood for a more radical transformation of economic relationships along corporatist lines.

On the whole, only from 1925 to 1934 did corporatist economics enjoy some popularity. Its partisans proclaimed that the *homo corporativus* should replace the individualist *homo oeconomicus*, but they failed to produce significant achievements in economic theory. Orthodox economists like Einaudi and Jannaccone, initially forced into a tactical retreat, took back the lead in debate after 1934. Most academic economists – Gustavo Del Vecchio (1883–1972), Marco Fanno (1878–1965), Costantino Bresciani Turroni (1882–1963), Giovanni Demaria (1899–1998), and others – put aside their laissez-faire beliefs and attempted to interpret corporatist economy from a marginalist viewpoint. Corporatism was thus reduced to a case of economic policy, which did not modify the content of pure theory. A characteristic that distinguished these and other economists was their firm attachment to

Paretian general equilibrium analysis, which they developed in a dynamic sense elaborating some suggestions derived from Pantaleoni's writings and from Pareto's sociology. At the same time, forced to defend orthodoxy against ideological attacks, these economists largely ignored or misunderstood the nature of the Keynesian revolution.

This success of orthodoxy can be mostly explained by institutional factors. The Fascist government tried to reform the organization of university studies, in 1935 transforming the teaching of economics into that of 'corporatist political economy'. However, orthodox economists jealously defended their academic autonomy, and the younger generation they recruited was composed of disciples whose career was generally not obstructed by political intrusions. The efforts of the Fascist regime concentrated on the creation of special schools and research institutions – such as the School of Corporatist Sciences of the University of Pisa, directed by Giuseppe Bottai (1896–1979), the Labour School of Florence, headed by Gino Arias (1879–1942), and the National Institute of Agrarian Economics, directed by Arrigo Serpieri (1877–1960).

Likewise, the major publishing houses – in particular Einaudi in Turin and Laterza in Bari – actively supported orthodox economics. The publisher that more actively sponsored corporatist economics was Sansoni in Milan, which issued a series connected to the Pisa Corporatist School, containing works on corporatism and economic planning. The lobbying activity of liberal economists also succeeded in modifying the editorial project of the *Nuova collana di economisti stranieri ed italiani* (1932–7), originally conceived as a sequel of the *Biblioteca dell'economista* and as the seal of Fascist economic culture. As a matter of fact this collection was open to recent international literature (Pigou, Sraffa, Hicks, Frisch, Hayek, Robertson and Keynes), and made no room for corporatist economics. Even the major cultural enterprise of the Fascist regime, the *Enciclopedia Italiana* edited by Giovanni Gentile, was quite impartial in the choice of authors for its economic entries.

Conversely, Mussolini's government was able to impose a considerable control over the periodical press. On the one hand, it created its own ideological mouthpieces – such as *Gerarchia* and *Critica Fascista* – and favoured the rise of economic journals – such as *Economia,* founded in 1923, and *Nuovi studi di diritto, economia e politica,* started in 1927 – that stimulated a considerable debate around the implications of corporatist economics. On the other hand, it extended its repression of journals of the liberal camp. Both *La Riforma sociale* and the *Giornale degli economisti* were discontinued for political reasons, in 1935 and 1942 respectively.

One of the costs of the Fascist years was a limited but significant 'brain drain': among those who were forced to emigrate were Bresciani Turroni, Umberto Ricci (1879–1946), Piero Sraffa (1898–1983), and the young Franco Modigliani (1918–2003).

The evolution of the economics profession in the post-war period was substantially influenced by the restoration of liberal-democratic institutions. First and foremost, the recovered political freedom favoured the rise of a network of centres of research and advanced studies (the Centre of Specialisation and Economic–Agrarian Research of Portici, the Svimez in Naples, the Istao in Ancona, the Research Department of the Bank of Italy in Rome) and of university departments.

Scholarships were granted to young scholars who wanted to continue their studies abroad, encouraging the opening of frontiers to international debate after the relative isolation of the Fascist period. The main economic journals were restructured and new specialized periodicals emerged, adopting international standards. Another crucial event was the creation in 1951 of the Società Italiana degli Economisti, whose constitution stipulated a full economics professorship as a criterion for admission.

The new political context soon stimulated many economists to return to their traditional public vocation. Einaudi, Labriola and Nitti sat in the Constitutional Assembly (1946) with other economists of the younger generation, including Epicarmo Corbino (1890–1984), Amintore Fanfani (1908–1999), Antonio Pesenti (1910–

1973), Paolo Emilio Taviani (1912–2001) and Ezio Vanoni (1903–1956). A special Commission on economic and social affairs nominated by the government was chaired by Demaria and composed of the most eminent amongst his colleagues.

It was mostly from the political side that Keynesianism made its entry into the Italian debate in the early 1950s, despite the persisting reluctance of economists to accept its theoretical underpinnings. Some Catholic economists engaged in politics, such as Fanfani and Giorgio La Pira (1904–77) declared themselves to have been inspired by Keynes when, as members of the cabinet, they introduced a plan for subsidized housing to reduce unemployment. But a Keynesian flavour could also be discerned in the 'Plan for labour' propounded in 1948 by the CGIL, the communist and socialist trade union. Likewise, the 'Scheme for the growth of employment and income in Italy in the decade 1955–1964' presented by Vanoni was explicitly inspired by Harrod's growth model. Finally, the debate of the 1960s on economic planning, in which Ferdinando di Fenizio (1906–74), Pasquale Saraceno (1903–1991), Giorgio Fuà (1919–2000), Paolo Sylos Labini (1920–2005) and Federico Caffè (1914–1987) participated, was clearly dominated by Keynesian assumptions. This does not mean that Keynes's theory was not present in more academic debates. The second edition of di Fenizio's *Lezioni di teoria economica* (1948) reflected the neoclassical synthesis arguing that the Keynesian approach was complementary rather than alternative to classical theory. Also Caffè and Vittorio Marrama (1914–82) published a series of theoretical contributions on Keynesian economic policies. Finally, Keynesianism exerted a considerable influence on the Italian public finance tradition, especially thanks to the works of Sergio Steve (1915–2006).

The 1960s were also marked by the impact on Italian economics of works of Sraffa, who since the 1920s had migrated to Cambridge. His article on 'The Laws of Return under Competitive Conditions' (1926) – preceded by a paper published in 1925 in the *Giornale degli economisti* – had criticized Marshall's equilibrium analysis and paved the way for research in imperfect competition. In *Production of Commodities by Means of Commodities* (1960) Sraffa expounded an alternative to general equilibrium analysis based on a reformulation of the classical and Marxian notion of surplus. This work originated a school of thought, the neo-Ricardians, that exerted a powerful influence on international scientific debates and on Italian academic life for a couple of decades. Among its main representatives one should count Pierangelo Garegnani and Luigi Pasinetti.

The strength of neo-Ricardian economics has probably been the last distinctive feature of Italian economics, at least in its most theoretical departments. In recent decades, the growing internationalization of this discipline has caused Italian economics to move towards the American-based mainstream of economics, with its typical formalistic and quantitative features. This change has been symbolized by the move to English as *lingua franca* not only of economic discussion but also of Italian journals, conferences and Ph.D. programmes.

See Also

- ▶ Barone, Enrico (1859–1924)
- ▶ Beccaria, Cesare Bonesana, Marchese di (1738–1794)
- ▶ Einaudi, Luigi (1874–1961)
- ▶ Galiani, Ferdinando (1728–1787)
- ▶ Genovesi, Antonio (1712–1769)
- ▶ Modigliani, Franco (1918–2003)
- ▶ Ortes, Giammaria (1713–1790)
- ▶ Pantaleoni, Maffeo (1857–1924)
- ▶ Pareto, Vilfredo (1848–1923)
- ▶ Sraffa, Piero (1898–1983)
- ▶ Verri, Pietro (1728–1797)

Bibliography

Asso, P.F., ed. 2001. *From economists to economists. The International Spread of Italian Economic Thought, 1750–1950.* Florence: Polistampa.

Augello, M.M., and M.E.L. Guidi 1996. The emergence of economic periodical literature in Italy (1750–1900).

In *Political Economy in European Periodicals, 1750–1900*, ed. M. Bianchini. Special issue of History of Economic Ideas 4(3), 15–62.

Augello, M.M., and M.E.L. Guidi. 2001. The associations of economists and the dissemination of political economy in Italy. In *The spread of political economy and the professionalisation of economists. Economic societies in Europe, America and Japan in the nineteenth century*, ed. M.M. Augello and M.E.L. Guidi. London: Routledge.

Augello, M.M., and M.E.L. Guidi. 2005a. The Italian economists in Parliament from 1860 to 1922: A quantitative analysis. *European Journal of the History of Economic Thought* 12: 279–319.

Augello, M.M., and M.E.L. Guidi. 2005b. Economists and political economy in Parliament from the unification of Italy to the rise of the Fascist regime (1861–1922). In *Economists in parliament in the liberal age, 1848–1920*, ed. M.M. Augello and M.E.L. Guidi. Aldershot: Ashgate.

Augello, M.M., and M.E.L. Guidi, eds. 2007. *L'economia divulgata (1840–1922). Stili e percorsi italiani*, *Vol. I*, *Manuali e trattati; Vol. II*, *Teorie e paradigmi; Vol. III, La «Biblioteca dell' economista» e la circolazione internazionale dei manuali.* Milan: Franco Angeli.

Bartoli, H. 2003. *Histoire de la pensée économique en Italie*. Paris: Publications de la Sorbonne.

Barucci, P. 1972. The spread of marginalism in Italy (1871–1890). *History of Political Economy* 4: 512–531.

Barucci, P., ed. 2003. *Le frontiere dell'economia politica. Gli economisti stranieri in Italia: dai mercantilisti a Keynes*. Florence: Polistampa.

Bellanca, N. 1993. *La teoria della finanza pubblica in Italia, 1883–1946. Saggio storico sulla scuola italiana di finanza pubblica*. Florence: Olschki.

Bianchini, M. 1982. *Alle origini della scienza economica: felicità pubblica e matematica sociale negli economisti italiani del Settecento. Parma: Studium. French translation: Bonheur public et méthode géométrique: enquête sur les économistes italiens (1711–1803)*. Paris: Ined, 2002.

Bianchini, M. 1989. Some fundamental aspects of Italian eighteenth-century economic thought. In *Perspectives on the history of economic thought*, ed. D.A. Walker. Aldershot: Edward Elgar.

Bianchini, M. 1994. The Galilean tradition and the origins of economic science in Italy. In *Political economy and national realities*, ed. M. Albertone and A. Masoero. Turin: Fondazione Luigi Einaudi.

Buchanan, J. 1960. 'La scienza delle finanze': The Italian tradition in fiscal theory. In *Fiscal theory and political economy. Selected esssays*. Chapel Hill: University of North Carolina Press.

Di Battista, F. 1983. *L'emergenza ottocentesca dell'economia politica a Napoli*. Bari: Facoltà di Economia e Commercio.

Faucci, R. 2000. *L'economia politica in Italia dal Cinquecento ai nostri giorni*. Turin: Utet Libreria.

Faucci, R., and S. Perri. 1995. Socialism and marginalism in Italy (1880–1910). In *Socialism and marginalism in economics, 1870–1930*, ed. I. Steedman. London: Routledge.

Fausto, D., and V. De Bonis, eds. 2003. The Italian Tradition of Public Finance, special issue of *Il Pensiero economico italiano* 11(1).

Garofalo, G. and A. Graziani, eds. 2004. *La formazione degli economisti in Italia (1950–1975).* Bologna: Il Mulino.

Guidi, M.E.L. 2000. Corporative economics and the Italian tradition of economic thought: A survey. *Storia del pensiero economico* 40: 31–58.

Mancini, O., F.D. Perillo, and E. Zagari, eds. 1982. *Teoria economica del corporativismo*. Vol. 2. Naples: Edizioni scientifiche italiane.

Meacci, F., ed. 1998. *Italian economists of the 20th century*. Cheltenham: Edward Elgar.

Nuccio, O. 1984–1987. *Il pensiero economico in Italia (1050–1750)*. Vol. 7. Rome: Mediocredito Centrale.

Parisi, D. 1984. *Il pensiero economico classico in Italia (1750–1860). Criteri definitori ed evoluzione storica*. Milan: Vita e Pensiero.

Porta, P.L. 1993. A note on Italian economics in the early nineteenth century from Restoration to Risorgimento. *History of Economic Ideas* 1: 43–70.

Porta, P.L. 1996. Italian economics through the postwar years. In *The post-1945 internationalization of economics.* Annual supplement to vol. 28 of *History of political economy.* Durham/London: Duke University Press.

Quadrio Curzio, A., ed. 1996. *Alle origini del pensiero economico in Italia. 2. Economia e istituzioni. Il paradigma lombardo tra i secoli XVIII e XIX*. Bologna: Il Mulino.

Roggi, P., ed. 2001. Le grandi 'voci' nei dizionari specializzati (e non) di economia. Special issue of *Storia del pensiero economico* 41–42.

Romani, R. 1994. *L'economia politica del Risorgimento italiano*. Turin: Bollati Boringhieri.

Roncaglia, A., ed. 1995. *Alle origini del pensiero economico in Italia. 1. Moneta e sviluppo negli economisti napoletani dei secoli XVII–XVIII*. Bologna: Il Mulino.

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