Technology and Finance

Challenges for financial markets, business strategies and policy makers

Edited by Morten Balling, Frank Lierman and Andy Mullineux

London and New York

Technology and Finance

Technology and Finance analyses the dramatic implications of technology for today's financial sector, for productivity growth and for monetary policy. A wide range of financial market activities are now technology driven; technology is also crucial in retail, private and corporate banking, and it has lowered entry barriers to the sector. New participants are flourishing as they are strongly supported by their technology.

Distinguished keynote speeches in this volume were initially featured in the latest highly-respected *SUERF* (*Société Universitaire Européenne de Recherches Financières*) Colloquium. The articles cover the following topics:

- The impact of technology on financial institutions – evolution or revolution?
- The relationship between technology and financial markets, including the impact of 'electronification' on financial markets and deposit insurance systems.
- The real and potential impact of technology on productivity growth, and possible implications for economic growth, monetary policy and markets.

This volume represents the cutting edge of informed thinking on the implications of, and possible problems with, modern technology on contemporary finance.

Morten Balling is Professor of Finance at the Aarhus School of Business, Denmark. He was president of the ASB from 1993–2001 and has been a council member of the *Société Universitaire Européenne de Recherches Financières (SUERF)* since 1994.

Frank Lierman is Chief Economist at DEXIA Bank, Belgium. A council member of *SUERF* since 2000, he is also president of the editorial board of the *Revue Bancaire et Financière/Bank – en Financiewezen*.

Andy Mullineux is Professor of Global Finance, and Director of the Global Finance Research Group in the Department of Accounting and Finance at the Birmingham Business School, University of Birmingham. His most recent publications include co-editing *Finance*, *Governance and Economic Performance in Pacific and South East Asia*, and *Economic Performance and Financial Sector Reform in Central and Eastern Europe*.

Routledge International Studies in Money and Banking

1 **Private Banking in Europe**
Lynn Bicker

2 **Bank Deregulation and Monetary Order**
George Selgin

3 **Money in Islam**
A study in Islamic political economy
Masudul Alam Choudhury

4 **The Future of European Financial Centres**
Kirsten Bindemann

5 **Payment Systems in Global Perspective**
Maxwell J Fry, Isaak Kilato, Sandra Roger, Krzysztof Senderowicz, David Sheppard, Francisco Solis and John Trundle

6 **What is Money?**
John Smithin

7 **Finance**
A characteristics approach
Edited by David Blake

8 **Organisational Change and Retail Finance**
An ethnographic perspective
Richard Harper, Dave Randall and Mark Rouncefield

9 **The History of the Bundesbank**
Lessons for the European Central Bank
Jakob de Haan

10 The Euro
A challenge and opportunity for financial markets
Published on behalf of Société Universitaire Européenne de Recherches Financières (SUERF)
Edited by Michael Artis, Axel Weber and Elizabeth Hennessy

11 Central Banking in Eastern Europe
Nigel Healey

12 Money, Credit and Prices Stability
Paul Dalziel

13 Monetary Policy, Capital Flows and Exchange Rates
Essays in memory of Maxwell Fry
Edited by William Allen and David Dickinson

14 Adapting to Financial Globalisation
Published on behalf of Société Universitaire Européenne de Recherches Financières (SUERF)
Edited by Morten Balling, Eduard H. Hochreiter and Elizabeth Hennessy

15 Monetary Macroeconomics
A new approach
Alvaro Cencini

16 Monetary Stability in Europe
Stefan Collignon

17 Technology and Finance
Challenges for financial markets, business strategies and policy makers
Published on behalf of Société Universitaire Européenne de Recherches Financières (SUERF)
Edited by Morten Balling, Frank Lierman and Andy Mullineux

Technology and Finance

Challenges for financial markets, business strategies and policy makers

Edited by Morten Balling, Frank Lierman and Andy Mullineux

London and New York

First published 2003
by Routledge
11 New Fetter Lane, London EC4P 4EE

Simultaneously published in the USA and Canada
by Routledge
29 West 35th Street, New York, NY 10001

Routledge is an imprint of the Taylor & Francis Group

This edition published in the Taylor & Francis e-Library, 2004.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Technology and finance: challenges for financial markets, business strategies, and policy makers / edited by Morten Balling, Frank Lierman, and Andrew Mullineux.
p. cm.– (Routledge international studies in money and banking ; 17)
Includes bibliographical references and index.

ISBN 0-203-22261-X Master e-book ISBN

ISBN 0-203-27707-4 (Adobe eReader Format)
ISBN 0-415-29827-X (Print Edition)

1. Banks and banking – Technological innovations–Europe. 2. Banks and banking–Technological innovations. 3. Financial institutions–Technological innovations–Europe. 4. Financial institutions–Technological innovations. 5. Financial services industry–Technological innovations. I. Balling, Morten. II. Lierman, Frank. III. Mullineux, A.W. IV. Series.

HG2974.T434 2002
332.1–dc21

2002026937

Contents

List of figures

List of tables

List of contributors

Helen Allen Adviser, Market Infrastructure Division, Bank of England, UK

Morten Balling Professor, Department of Finance, Aarhus School of Business, Denmark

Cláudia Costa Storti Economist, Banco de Portugal, Lisboa, Portugal

Jon Danielsson Lecturer in Finance, Financial Markets Group, London School of Economics, UK

Paul De Grauwe Professor, Centrum voor Economische Studiën, Katholieke Universiteit Leuven, Belgium

Hans Degryse Economics Department, KU Leuvew, Belgium

Marieke Donker Researcher, De Nederlandsche Bank, Amsterdam, The Netherlands

Ignacio Fuentes Egusquiza Senior Analyst, Servicio de Estudios, Banco de España, Madrid, Spain

Charles Goodhart Professor, Banking and Finance, London School of Economics, UK

Michael H. Grote Lecturer, Institute for Economics, Goethe-University, Frankfurt/Main, Germany

John Hawkins Senior Economist, Bank for International Settlements, Basel, Switzerland

Barry Howcroft Professor, Director, Banking Centre, Loughborough University, UK

Olivier Lefebvre Executive Vice-President, Member of the Managing Board, Euronext NV, Brussels, Belgium

Gottfried Leibbrandt McKinsey and Company, Amsterdam, The Netherlands

Frank Lierman Chief Economist, DEXIA Bank NV, Brussels, Belgium

Kjersti-Gro Lindquist Senior Advisor, Research Department, Central Bank of Norway, Oslo, Norway

David T. Llewellyn Professor, Department of Economics, Loughborough University, UK

Vivien Lo Research Assistant/Lecturer, Institute for Economic and Social Geography, Goethe-University, Frankfurt/Main, Germany

Andy Mullineux Professor of Global Finance, Department of Accounting and Finance, University of Birmingham, UK

Ralf-Henning Peters Assistant Professor, Department of Economics, Otto-von-Guericke, University of Magdeburg, Germany

Guy Quaden Governor, National Bank of Belgium, Brussels, Belgium

Didier Reynders Minister of Finance, Brussels, Belgium

Teresa Sastre de Miguel Senior Economist, Servicio de Estudios, Banco de España, Madrid, Spain

Setsuya Sato Head of Special Meetings, Bank for International Settlements, Basel, Switzerland

Jean-Paul Servais Deputy Director of Cabinet, Ministry of Finance, Brussels, Belgium

Peter D. Spencer Professor, Economics Department, Birkbeck College, London, UK

Antje Stobbe Head of e-Research, DB Research, Deutsche Bank AG, Frankfurt/Main, Germany

Siegfried Utzig Director Economics, Capital Markets, Bundersverband Deutscher Banken, Berlin, Germany

Mark Van Achter Economics Department, Katholieke Universiteit Leuven, Belgium

Johan Van Gompel Head of Economic Research, KBC Asset Management, Brussels, Belgium

Iman Van Lelyveld Researcher, De Nederlandsche Bank, Amsterdam, The Netherlands

Peter Westerheide Senior Researcher, Centre for European Economic Research, Mannheim, Germany

Acknowledgements

The editors and the publishers gratefully acknowledge permission to reproduce copyright material from the following:

Nomos-Verlag, Baden-Baden, for permission to reprint several graphs from volume 58 of the *ZEW-Wirtschaftsanalysen* in Chapter 7 of this volume: ‘Employment perspectives in the German financial services industry and the impact of information technology’ by Ralf-Henning Peters and Peter Westerheide.

Ralf-Henning Peters acknowledges financial support by Deutsche Ausgleichsbank Bonn.

Journal of Financial Regulation and Compliance Volume 10, Number 1 (2002) for permission to reprint Guy Quaden, ‘Central banking in an evolving environment’ as Chapter 1 of this volume. This article is reproduced with the permission of Henry Stewart Publications (HSP), 25 Museum Street, London WC1A 1JT.

Every attempt has been made to obtain permission to reproduce copyright material. If any proper acknowledgement has not been made, we would invite copyright holders to inform us of this oversight.

Introduction

Morten Balling

The papers in this volume were presented at a colloquium jointly organized in Brussels 25–27 October 2001 by the *Société Universitaire Européenne de Recherches Financières (SUERF)* and the *Belgian Financial Forum (BFF)*. The theme of the colloquium was 'Technology and Finance: Challenges for Financial Markets, Business Strategies and Policy Makers'. The authors illuminate a large number of important aspects of this theme. Technology has important implications for the earnings, costs, risks, competitiveness and location of financial institutions. It affects the way securities transactions are carried out, the transparency of the markets, settlement activities and the structure of the exchange industry. Technology changes payments systems and the framework for strategic decisions in the financial industry and for monetary policy. Financial supervision and regulation must be adapted to new risks and new risk management methods.

The book is organized in five parts. Part I consists of three lectures from plenary sessions, Part II of contributions on technology and financial institutions, Parts III and IV contain respectively contributions on technology and financial markets and on technology and payments. Part V deals with technology and productivity.

In the first chapter in Part I, *Guy Quaden* (Governor of the Bank of Belgium) analyses the changes caused by technological innovation to the macroeconomic environment and in the financial sector. Integration of new technology and deregulation have both contributed to a reshaping of the financial landscape. A more open, competitive and globalized financial market has emerged. Not all financial institutions have understood how to take the new uncertainties and risks for financial sectors into account. Central banks have to provide monetary stability. Maintaining price stability is the primary objective of monetary policy. The development of e-money is now so important that e-money issuers should not escape reserve requirements. Technological and financial market changes make it necessary for central banks to reassess the information content of many economic indicators. An increasing number of central banks now complement their traditional annual reports centred on monetary policy and macroeconomic developments by another report focused on the theme of financial stability. New technology is also playing a crucial role in the development of secure and efficient payment and settlement systems. Oversight of these systems has become a key function in modern central banking. Finally, the more globalized financial market calls for a more globalized approach to supervision. The monetary stability and the financial stability wings belong to the same bird.

In the second chapter, Professor *Charles Goodhart* and *Jon Danielsson* look at risk from a time perspective. Risk must be understood in the context of a shifting and unpredictable world. Credit risk of financial institutions and the need for provisioning for bad debts are affected by largely unpredictable business cycles. Indeed the term ‘cycle’ may be a misnomer. Technology has an important impact on trends in productivity, asset price movements and economic activity. It is, however, very difficult for private market participants and for financial regulators to establish at which stage in the so-called cycle they are and when the next recession will come. Historically, there is very little sign of any constant regularity, or periodicity, in the onset of recessions. It is equally difficult for regulators or anyone else to determine quantitatively the extent of asset price misalignment. Hence, regulators should – in spite of these macroeconomic measurement problems – try to develop a system with counter-cyclical movements in the regulatory variables such as collateral requirements, loan-to-value ratios and minimum capital requirements.

In Chapter 3, Minister of Finance *Didier Reynders* and Deputy Director of Cabinet *Jean-Paul Servais* look at the new challenges for European and national regulations relating to financial markets. The Belgian presidency of the European Council has decided to continue its work for the implementation of the proposals in the Lamfalussy Report and the drafts concerning the prospectus directive, the prudential supervision directive, the insider trading directive and the draft regulation concerning mandatory use of IAS – International Accounting Standards, at least as far as consolidated accounts are concerned. The proposed directive on collateral and financial guarantees is considered important in derivatives markets. Technological innovation has induced the development of a new licensing system for settlement and clearing institutions and formulation of rules concerning alternative trading systems. Finally, the Belgian presidency is in charge of preparing a draft directive concerning money laundering, considered as urgent since the terrorist attack of 11 September 2001.

Part II consists of Chapters 4 to 8 and deals with the impact of technology on financial institutions.

In Chapter 4, Professor *David T. Llewellyn* argues that the new technology causes changes in the underlying economics of banking in a rather fundamental way. New types of competitors enter the market for banking services. Technology affects entry conditions, management methods, production processes and distribution channels. In response, banks will have to adjust the way they conduct their business, to make their delivery systems more efficient and to revise their organizational structures. Appropriate responses can also be formation of joint ventures, outsourcing and subcontracting of some services. One of the implications of the new technology is that it has become unclear what the optimal financial organisation and structure are. Plurality can be expected in the financial industry. Small banks are not condemned to disappear. Technology offers alternatives to realize economies of scale and creates the possibility to unbundle processes.

In Chapter 5, *Ignacio Fuentes Egusquiza* and *Teresa Sastre* study the effect of technology on the costs and risks of Spanish banks. The spending on information technology and telecommunications had a slightly different time pattern in commercial banks, savings banks and credit cooperatives respectively. Due to their strong involvement in the interbank and government bonds markets, the Spanish commercial banks had to adapt quickly to the new technical possibilities and to the requirements derived from the implementation of a single currency. Investment in new technology and a trend towards remote banking reduces the weight of personnel costs in total operating costs of the commercial banks. Cost patterns in savings banks and credit cooperatives have been somewhat different. These institutions have in fact increased their branch networks and the number of employees has grown. Spanish banks have worked for some time to develop and implement new risk control systems. The new risks arising from electronic banking is taken into consideration. This work will continue in cooperation with the Spanish authorities.

In Chapter 6, Professor *Barry Howcroft* investigates consumer behaviour in connection with remote and direct banking. He applies a two-dimensional matrix of consumer behaviour in which the consumer's confidence and involvement are cross-classified. The data from a survey based on a questionnaire are analysed with application of cluster analysis. The respondent customers are grouped in six clusters with different degrees of confidence in their financial institution and different involvement in financial transactions. An interesting result is that bank customers with a relatively low income tend to prefer to use branch networks rather than technology-driven and remote delivery channels. This suggests that the least profitable segments of the bank's customer base have a predisposition to use the most expensive delivery channel. In contrast, the most efficient and profitable customers are predominantly using the most cost effective delivery channels.

In Chapter 7, *Ralf-Henning Peters* and *Peter Westerheide* assess the determinants and perspectives of employment in Germany's financial services industry. In 1997, about 1.25 million people worked in the industry. Their analysis shows that employment in the financial services industry can be expected to decrease in the coming years. The decrease will in particular affect employees with a low degree of education. For highly qualified personnel, growing employment can be expected. In an econometric time series analysis it is demonstrated that the different skill groups have been influenced in different ways by increasing use of information and communication technology. The results correspond with the results of a survey among financial market experts.

In Chapter 8, *Iman van Lelyveld* and *Marieke Donker* study the effects of new information and communication technology on the location of financial activity. They investigate what they call the 'Geography Doesn't Matter' hypothesis, which says that the production of financial services could take place anywhere and that the easier and cheaper communication means that geography has become irrelevant. Financial activity is measured at the regional level as gross value added and employment in the financial sector. If the hypothesis is true, the distribution of financial activity should become more even across regions during the period

(1980–95) covered by the data. The authors find that, with some minor exceptions, there have not yet been large shifts in the regional distribution of production and employment in financial services in Europe.

Part III consists of Chapters 9 to 12 which consider the impact of technological change on financial markets.

In Chapter 9, *Olivier Lefebvre* draws on his experiences from Euronext and presents his views on competition and consolidation in the European exchange industry. National exchanges compete in the primary market for listings, but the companies have so far been most interested in listing on the home market exchange. In the secondary market, exchanges compete for liquidity. Since the direct exchange cost represents a relatively small part of total transaction cost, it is difficult for the individual exchanges to attract trading activity from other exchanges. In addition, the relative performance of the national exchanges is primarily explained by the performance of their national blue chips and not by the competitiveness of the exchanges as such. In-house matching between the customers of big banks is potentially the most serious threat to the organized exchanges in Europe. Euronext recommends the establishment of a regulatory environment in Europe with true pan-European exchanges and a level playing field as objectives. There should be a ‘single passport’ for exchanges giving them access to all fifteen jurisdictions. Regulation should be carried out ‘by function’. All kinds of order matching services including in-house matching should be regulated according to the same rules. A regulatory environment according to these principles will provide the best framework for investor protection.

In Chapter 10, *Hans Degryse* and *Mark Van Achter* analyse the impact of alternative trading systems on transaction costs and liquidity. Their paper was awarded the Marjolin Prize 2001. Alternative trading systems are defined as trading mechanisms developed independently from the established marketplaces and designed to match buyers and sellers on an agency basis. In the USA, such systems have been successful in attracting trade, in particular in the NASDAQ dealer market. Electronic communication networks have helped to lower trading costs significantly. They have also improved the quality of markets. Alternative trading systems have not been as successful in European financial markets. The authors explain this by referring to the fact that traditional exchanges in Europe have created efficient electronic trading facilities themselves earlier than their US counterparts. They mention also that several European markets are organized as auction markets.

In Chapter 11, *Michael H. Grote* and *Vivien Lo* look at the geographical location of stock market trading. Network externalities are remarkable features of stock exchanges. For each actor the utility of using the market increases when new participants join the market. The larger the market share of a stock exchange, the higher the probability that a new user will choose it. Consequently, network externalities and liquidity considerations have provided strong reasons for traders to concentrate spatially, close to stock exchanges and financial centres. The authors call this ‘agglomeration effects’. The implementation of new information and communication technologies transform these centripetal forces. Spatial proximity to the market can be substituted by virtual proximity. Agglomeration of traders and exchange of market information can take place within electronic networks. There is, however, an important difference between straightforward information, which is readily transferable via telecommunication, and complex information that is not. Location of traders has therefore not become irrelevant. Face-to-face contact with analysts has value. Closeness to other traders and company insiders is important. Informal meetings can increase understanding. It is consequently not easy to project where traders will go when stock exchanges go virtual.

In Chapter 12, *John Hawkins*, *Helen Allen* and *Setsuya Sato* consider the transformation of the economic landscape of trading venues caused by the adoption of electronic trading systems. The new systems have implications for the architecture and quality of financial markets. Ultimately, there are broader welfare implications related to efficiency and financial stability. There are multiple, possibly competing, public policy objectives along with uncertainties about the net effect of changes in markets and their transmission to broader welfare. Regulatory agendas all over the world deal with issues related to electronic trading. It must for instance be decided whether or not the frameworks for regulation should continue to differentiate between exchanges and non-exchange trading systems. Another important issue is the degree of detail and enforcement in transparency rules. Authorities responsible for financial system stability should adapt their policies to the fact that new and different firms are involved in the operations of financial markets with possible implications for systemic risks.

Part IV consists of Chapters 13 to 16 which analyse the impact of technological change on payments systems.

In Chapter 13, *Cláudia Costa Storti* and *Paul De Grauwe* analyse the implications of a cashless society for the conduct of monetary policy. They try to establish whether there will be a mechanism that ties down the price level and prevents systematic inflation. Furthermore, they analyse whether and how the central bank can be transformed so that it can maintain price stability in a cashless society. They conclude that the central bank will lose its traditional instruments of monetary policy. It might be relevant to redefine the role of the central bank. Its supervisory role could be strengthened. The quality of the loan portfolios of private money-issuing institutions could be controlled by the central bank. This kind of supervision should also be expanded to issuers of e-money.

In Chapter 14, *Gottfried Leibbrandt* studies trends in the use of payment instruments. There are still large differences among countries between the relative roles of automated clearing, giro, credit cards, debit cards and cheques, respectively. The differences have to a large extent historical reasons. Electronic payment instruments are significantly cheaper than paper-based instruments. The fee structure has a strong impact on the customers' choice of payment instrument. In most of continental Europe, banks have a long tradition for cooperation in matters of payment networks. The author tries to evaluate whether Europe will adopt a single compatible giro- and card-system. His expectation is that differences between payment systems will persist for some years. A pan-European payment network may be further off than most people think. In spite of the introduction of the euro, for the time being 'one size will not fit all'.

In Chapter 15, *Kjersti-Gro Lindquist* analyses the importance of new payment systems for the development in banks' scale properties and input demand in their production of loans and deposits. The author estimates a four-factor translog cost function and estimates the corresponding cost-share equations using Norwegian unbalanced bank-level panel data for the period 1987–98. The inputs are labour, physical capital, materials and funding. The sum of loans and deposits is treated as an aggregate output. The results show that the increase in electronic payments has increased the elasticity of scale and decreased average variable costs in banking. The move towards electronic payment systems has affected input demand asymmetrically, i.e. non-neutral, causing the cost-shares of both labour, physical capital and materials to increase, while the cost-share of funding has decreased. New electronic payment systems have particularly substituted out paper-based and labour-intensive methods.

In Chapter 16, *Peter D. Spencer* tries to answer the question: Why have digital cash systems failed to penetrate the payments market while electronic trading of securities has been a success? The author compares different payment instruments with respect to their costs and risks. The convergence of Internet, television and telephone systems means, according to the author, that there is a huge digital market to be exploited. Market contestability and customer switching costs play a significant role. Presently, entry, exit and switching costs are high in markets for payment systems. He criticizes the banks' fee structures for being out of line with the cost structure. Zero-interest transaction accounts with low transaction charges are common. In the author's view, regulators should pay careful attention to the pricing of transaction media. The regulator should also try to ensure open markets and minimization of switching costs in order to foster the development of digital money.

Part V consists of Chapters 17 to 19 which investigate the impact of technological change on productivity.

In Chapter 17, *Antje Stobbe* looks at the 'new economy' in a European context. Different studies of US data have tried to explain the growth of labour productivity partly as a reflection of spillover effects from increasing use of information and communication technology. The effect of higher investment in the new technology on total factor productivity is, however, uncertain. The author uses household data in order to illuminate the degree to which diffusion of information and communication technology varies from country to country within Europe. The evidence shows that use of information technology correlates positively with GDP *per capita*. PC and Internet penetration is higher in the Nordic countries than in the southern European region. An empirical study from ECB based on European data suggests that positive spillover effects from the use of information and communication technology have only been limited if present at all. Significant new economy effects cannot be observed for the EU as a whole. So far the 'new economy' in Europe is more mirage than reality.

In Chapter 18, *Siegfried Utzig* evaluates the role of innovations in information technology for growth in productivity. Researchers have to deal with difficult measurement problems. How should one define output which involves a quality dimension without reference to innovation? What exactly are the units of output in banking? Unless the nature of the output can be defined precisely, it is impossible to determine its rate of growth or to answer questions about the impact of quality-enhancing innovations like those of communications technology. The success of the new economy can only be assured by economic policy. It should create entrepreneurial room and avoid conservation of existing structures that tend to obstruct the transition to the new economy. In the USA confidence in market forces led to deregulation in telecommunications and the flourishing of communications technology companies. Europe lags behind in this area. The new economy does, however, not require a new economic policy. Both overestimation and underestimation of productivity growth can lead to mistakes in economic policy. What is really important, however, is to improve the general conditions for innovation and entrepreneurship so that they can meet the challenges posed by the technological transition to the information society.

In Chapter 19, *Johan Van Gompel* discusses the extent to which the utilization of technological progress in the production process is influenced by the structural-institutional characteristics of a country and which policy measures the government can take (or must avoid) to allow innovating economic activities to develop to the full. He looks at productivity trends since 1960. In spite of a massive increase in investment in the application of new technologies, the mean annual growth of total factor productivity fell in the OECD area from 1960–73 to 1973–95. This so-called ‘Solow Paradox’ has several explanations. One possible explanation is that it takes a considerable amount of time before new computer technologies are fully disseminated. Other possible explanations have been given: higher tax rates, stricter environmental regulations, and the development of social security systems have led to a more unfavourable business climate. There are significant differences among countries. In the 1990s, the growth in productivity was relatively rapid in the USA. Some researchers explain this with a high level of investment in information and communications technology, while other researchers estimate that a considerable part of the growth since 1995 is of a cyclical nature. The flexibility of the labour market is also important for macroeconomic performance. The EMU countries tend to lag behind the USA in terms of dynamism and technological innovation.

Part I

Survey lectures from plenary sessions

1 Central banking in an evolving environment*

Guy Quaden

Ladies and gentlemen,

I am particularly pleased to speak before this eminent forum and to have the opportunity of addressing the topic 'Technology and Finance' from a central banker's point of view.

Technology and the challenges it raises for financial markets is a most appropriate theme for this colloquium. On the one hand, technology is closely intertwined with the evolution of many other factors affecting financial markets and so allows coverage of a wide range of issues. On the other hand, it has far-reaching consequences for all market participants, and this certainly includes central banks. In the first part I propose to briefly recall how technology has impacted on our macroeconomic and regulatory environment. In the next two parts I would like to sketch the main consequences of these developments on what my colleagues at the Bank of England have described as the two wings of central banking, i.e. monetary and financial stability.

Technology is a powerful factor of change in our *macroeconomic environment*. So, a few years ago, our colleagues of the Federal Reserve had to recognize that something new was happening in the US economy: a persistent higher growth and lower unemployment without the emergence of inflationary strains. This was related, at least partly, to the revolution in the information and communication technologies, which increased productivity growth and fostered efficiency in the labour market too. A third feature of the so-called new economy, the reduction in the variability of output growth, obviously proved to be short-lived! But a wave of over-pessimism should not submerge the previous wave of over-optimism. The question whether the American economy is still on a higher trend productivity growth path remains open, as well as what the prospects for the European economy are in this respect. Europe will benefit from a specific driving force, the completion of the single market with the new single currency, which should trigger further structural reforms and hence foster innovative energies.

Technology also radically transformed the *financial sector*, which by the way greatly contributed to the new technological wave by financing it. New techniques

* Keynote speech given at 23rd SUERF Colloquium, Brussels, 27 October 2001.

in the treatment, the storage and the transfer of information exerted profound effects on a sector which is largely an information-based industry.

In a first stage, information technologies made it possible to develop more sophisticated products, to build up a better market infrastructure, to implement more accurate and reliable techniques for the control of risks, to reach more distant and diversified markets, and to multiply the value and the volume of operations. In short, new technology has radically transformed all three major functions performed by banks, i.e. access to liquidity, transformation of assets and monitoring of risks.

A new phase is presently at work with the emergence of e-money, e-banking and e-finance. It is clearly this new development which will represent the great challenge of the coming years. The speed of adoption of these new products remains difficult to forecast. Contrary to the preceding phase, this new wave is not limited to professional operators but involves all customers, including the retail market. Many of the scenarios suggested by IT firms or consulting groups have proved to be overly optimistic. At the same time, it would be wrong to become complacent. Most new technology is spreading following an S-shaped curve. The base section of the S can be quite long and practically horizontal; however, it will sooner or later be succeeded by a steep section. The example of Nordic countries, and more specifically Finland, shows how quickly e-finance can develop, once circumstances are ripe.

It is important to emphasize that the integration of new technology into the financial sector did not take place in isolation. Rather, it is the *interaction of technology* with another major development, *deregulation*, that contributed to reshaping the financial landscape. True, the pressure of the market to fully exploit the new technologies was strong, probably even irresistible. However a receptive environment had also been created by a lifting of the rather strict financial and banking regulations which were still prevalent at the end of the 1970s.

The increased awareness of the wide-ranging possibilities offered by new technologies illustrated that a lot could be gained through the removal of distortions in competition, directly linked to excessive regulation or intrusion from the government.

In combination, these two evolutions contributed to the emergence of a more open, competitive and globalized financial market, which obviously improves efficiency in the world economy. The transition, however, has not been a smooth one. It was not easy for the authorities who had to cope with the more frequent arbitrages operated by market participants between the various currencies and financial instruments or even between different legal, regulatory or tax regimes. Neither was it easy for financial intermediaries that had to work in a much more competitive environment where traditional protection and barriers to entry were progressively lifted. In short, the shift toward a more liberalized system together with the quick expansion of new products and markets has greatly increased uncertainties and risks for financial sectors.

This was not immediately recognized by many market participants who had previously been sheltered by the existing regulation. As a result, individual bank failures and banking crises, quasi non-existent between the end of the 1940s and the early 1970s, became all too frequent during the past two decades. In this context, the absence of significant problems within the Belgian banking sector during the recent period must be considered more an exception than a rule.

Despite deregulation there is still a major role to be played by *public authorities*, among others in the field of competition rules, consumer protection, fight against money laundering – and, of course, central banking. Central banks indeed have to provide stability, which certainly does not mean 'no change' but, on the contrary, building the best foundations for a sustainable dynamism. In doing so, central banks will cope with the evolving environment shaped by technical progress, by deregulation and globalization and, last but not least in Europe, by the single currency.

As regards the first wing of central banking, *monetary stability*, central banks have to provide a durable anchor in order for the price system to appropriately guide economic decisions. A stable value of money is all the more necessary for preserving the information value of relative prices in a changing world, where decisions have to be taken rapidly. Maintaining price stability is the primary objective of monetary policy, not only for the Eurosystem – according to the Treaty of Maastricht – but also for every central bank.

Nowadays the only regulations central banks rely on in designing the *operational* framework of monetary policy are the monopoly of banknote issuance and reserve requirements. The Eurosystem fully respects the principle of an open market economy with free competition, as enshrined in the Treaty. Its main instrument is the weekly allotment of credit by euro area-wide tenders. Minimum reserves, which are remunerated, have a stabilization function, thanks to an averaging provision, and are enlarging the structural liquidity shortage of the money market.

As the development of e-money is liable to weaken the leverage of the Eurosystem and in order to provide for a level playing field, e-money issuers should not escape reserve requirements. A European directive of last year rightly broadens the definition of credit institutions in order to include e-money institutions.

Technological change and financial market developments do not only affect monetary policy instruments but also the whole transmission process and consequently the *strategy* of monetary policy. In this complex and changing world the Eurosystem was right in rejecting any simple rule and adopting an all-encompassing two-pillar strategy. Central bankers have to continuously reassess the information content of many economic indicators. Let me pick out some of them – output, money, stock prices and bond market indicators – not because other variables, like wage developments and the fiscal policy stance, are less important, but because the former are most affected by technological and financial market changes.

Central bankers, even in the Governing Council of the ECB, are not insensitive to *growth* and employment prospects, as some critics argue. But they are well aware of two limitations: first, growth should not be stimulated to the detriment of price stability, because such a stimulus would be short-lived and would imply longer-term costs; second, as '*à la plus belle fille du monde on ne peut demander* *que ce qu'elle a*', monetary policy may exert some influence on the demand side of the economy but cannot solve structural problems, like persistent unemployment. Central banks may nevertheless contribute to output stabilization, as far as the risks to price stability are linked to the business cycle.

A central concept in this respect is the output gap, but its measurement, especially in real time, is surrounded with a large degree of uncertainty. Potential output growth, which is an ingredient of both pillars of the Eurosystem's strategy, is not known with precision. Should the new economy materialize, higher rates of growth could become sustainable. In the absence of any firm evidence of a new economy in the euro area – although some driving forces are to some extent in place – and since the emergence of a new economy is not driven by monetary policy, the Eurosystem did not take the risk of pre-emptively accommodating it. Nevertheless it monitors a wide range of indicators in order to periodically reassess the 'speed limit' of the euro area economy.

Needless to say, in the present circumstances growth is unfortunately even below the old economy speed limit, and the associated decrease in inflationary pressures has already prompted a 100 basis points interest rate cut in three steps since the spring.

The first pillar of the strategy of the ECB gives a prominent role to *money*. It is based on the conviction that inflation is a monetary phenomenon in the long run and underlines the medium-term orientation of monetary policy and the inheritance in this respect from the Deutsche Bundesbank. Recognizing that the demand for money can be subject to short-term fluctuations which are harmless for price stability, the ECB has not announced an 'intermediate objective' but rather a 'reference value' for the growth of a broad monetary aggregate. The recent rise in M3 growth is up to now interpreted as being such a short-term fluctuation, caused by the relatively flat yield curve and the weakness in stock markets.

Technology and financial market changes obviously affect the first pillar. They might increase the volatility of the income velocity of monetary aggregates and, as they are blurring the frontiers of 'moneyness', they complicate the definition of key aggregates. To paraphrase a former Governor of the Bank of Canada speaking about M1 twenty years ago, I would say that while the ECB is not planning to abandon M3, I cannot rule out that, some day, M3 could abandon us, but you already noticed that the first pillar is much more than the reference value. It encompasses a broad monetary analysis which will duly take into account such developments.

The first pillar also rests on the fact that credit institutions remain major players in the transmission process of monetary policy. The development of euro area capital markets could however increase the weight of financial market indicators in the second pillar.

The *stock market* is still a much less important channel of transmission in the euro area than in the USA, but the holding of shares is spreading, for example through mutual funds. I would not attempt to summarize the vast debate about the appropriate monetary policy reaction to asset price movements. I am inclined to say that central banks have not to put on these variables more emphasis than warranted by their effects on demand and should avoid asymmetric reactions –

benign neglect in the case of irrational exuberance, intervention in the case of sharp downward correction – which could pose a moral hazard problem.

The *bond market* provides indicators which are probably more important within the second pillar of the Eurosystem's monetary policy. Technical progress and European integration lead to more sophisticated and liquid markets which supply useful information about market expectations. Incidentally I notice that, while many central banks looked disapprovingly on indexed bonds prior to monetary union, the Eurosystem now welcomes the opportunity to extract information on inflation expectations from the comparison of yields on indexed and nominal bonds. Despite the upsurge in inflation in the euro area resulting from oil price and food price shocks, inflation expectations appear to remain very moderate, showing that the Eurosystem benefits from a high degree of credibility. Such a capital of credibility has to be preserved.

About the second wing of central banking, the safeguarding of *financial stability*, I would like to adopt a chronological approach. First, how are central banks currently adapting to the new environment by reconsidering the role they are playing in the financial market? Second, how could new technology affect the relations in the coming years between market participants, central banks and other supervisory and regulatory authorities?

Financial market developments and the heightened risks associated with these rapid changes led central banks to *reconsider the role they had to play* to preserve financial stability. For those central banks that were in charge of the surveillance of individual credit institutions, the implications were straightforward. They had to adapt the modalities of their microprudential activities. However, the need to proceed to *macroprudential monitoring* was also strongly felt by central banks, like the NBB, which were not vested with the microsupervision.

First, at an *analytical level*, central banks were induced to enlarge the scope of their research. The use of new technologies has caused a spectacular expansion in the volume of financial operations, certainly in comparison to the growth of real activities. This has required reconsidering the direction of the links between these two fields. Central banks had traditionally focused on the consequences that changes in financial conditions could have on the real economy. If such analyses remain essential, central banks are also increasingly concerned by the vulnerability of the financial system to fluctuations in real activities. So, the causalities also have to be reversed and due attention must be given to the impact that evolution of the real economy could have on the stability of the financial system. It is no coincidence that an increasing number of central banks now complement their traditional annual reports centred on monetary policy and macroeconomic developments by another report focused on the theme of financial stability. This is a development that the NBB will also actively embrace through the publication, possibly starting in 2002, of a new yearly Financial Stability Report.

At a more *operational level*, central banks contribute directly to strengthening the stability of the financial system by the development of secure and efficient payment and settlement systems. Here also new technology is playing a crucial role. Real time gross settlement systems, delivery versus payment mechanisms, cross-border connections between various clearing or settlement institutions, instant world transmission of information would be in practice unmanageable without the possibilities offered by IT technologies. These multiple layers of networks are too often considered as mere plumbing. However, this so-called plumbing is in many respects as spectacular and sophisticated as the more glamorous Internet or mobile phone networks.

The oversight of these modern payment and settlement systems has become a key function in modern central banking and this certainly applies to the NBB, as Belgium is hosting two major international institutions, SWIFT and Euroclear.

The second step in our chronological approach is also the most uncertain as it implies speculating about the impact of new technology on the *future organization* of financial markets.

It must be recalled from the outset that the introduction of new technology in the banking sector is not a one-shot phenomenon. On the contrary, it is proceeding by *waves*. As already said, the development of e-money, e-banking and e-finance will represent a great challenge.

Whatever its speed, this new wave will strongly modify the nature of relations between market participants. Distant access to financial products is substituting for close individual contacts. Brand loyalty, while still a key asset in a business built on trust, is increasingly associated with cherry picking. Banks themselves tend to shift from an approach based on long-term and stable relations to a strategy where each deal is individually appreciated on its own merits.

The various financial institutions are also redefining what should be their core business. The technological wave of the 1980s and early 1990s allowed the unbundling of most financial products into their various components. To the unbundling of products is now associated, thanks to the second wave of innovation, an outsourcing of the production and also of the distribution process. Back office functions, distribution networks and IT infrastructures can now easily be subcontracted, creating a new web of connections between various categories of market participants.

There are also important changes in the *relations between monetary and prudential authorities*, on the one hand, and *financial institutions* on the other. First the authorities will have to rely, much more than in the past, on the markets themselves for the surveillance of financial stability. One may legitimately feel concerned by such an evolution, which sounds like asking the fox to watch over the hens. However we must realize that financial markets are not only a major factor of change, they are also potentially a powerful factor of discipline. They are forcing credit institutions to be more transparent and to communicate more reliable information. The development of new, more sophisticated, risk management techniques, under the form of internal models, has been, at its roots, a private initiative from market participants. In order to integrate this modelling approach in the monitoring of banks' solvency, the Basel Committee on Banking Supervision is not designing a new system from scratch. On the contrary, it is referring to the best practices of the market itself as a benchmark against which to calibrate its own proposals.

Now, the authorities should not delude themselves. Best practices are what they are, easier to respect by strong institutions and in favourable circumstances, but much harder to maintain when the situation deteriorates. Far from being lightened, the burden of prudential authorities is becoming heavier. The necessarily limited human and financial resources of these authorities will be called upon all the more by the new Basel proposals which will require a more individualized and detailed surveillance of these new internal risk management systems.

The resource constraint will finally also strongly determine the *relations* that the *various monetary and supervisory authorities* will have to maintain *among themselves*. A more globalized financial market calls for a more globalized approach to supervision.

At the international level, several cooperative bodies and mechanisms have been established, either in the form of multilateral forums bringing together the competent authorities in the fields of prudential control and financial stability, or by means of bilateral protocols concluded between the supervisory bodies of different countries or sectors.

At each national level, the authorities also have to carry out an in-depth examination of their supervisory structure and procedures. I will not dwell here on the subject of the devolution of prudential tasks. Different models exist across the world bearing witness to the trade-offs which have to be made to adapt to market trends while also taking account of the specific national context.

Finally, this existing national and international framework needs to be periodically reviewed and adjusted. Whatever its form, the prevailing structure will have to fulfil two major conditions. On one hand, it must be efficient in preventing either loopholes or redundancy in supervision. On the other hand, it must be all-encompassing by combining the microprudential control of individual institutions with macroprudential monitoring of the systemic risks faced by the global financial market.

To conclude, let me stress that central banks are fully aware of the close connection and the large convergence existing between the two goals of *financial stability* and *monetary stability*. Keeping inflation under control, which is the ultimate goal of every central bank, has proved to be the best way to reduce uncertainties on the market, to alleviate distortions and, so, to eliminate one of the fundamental sources of financial instability.

Conversely, central banks need sound and efficient banking systems for ensuring rapid transmission, to the whole economy, of the impulses of their monetary policy. This is all the more important given that the assets at the disposal of central banks – the monetary base in our jargon – is becoming increasingly tiny compared to the total assets managed by credit institutions and, beyond that, by financial market operators.

In this context, the monitoring of financial stability may certainly not be considered as a by-product or a mere extension of the traditional monetary stability objective of central banks. The two functions are closely related but distinct. In other words, the monetary stability and financial stability wings belong to the same bird.

2 The inter-temporal nature of risk*

Charles Goodhart and Jon Danielsson

Risk arises primarily from two essential, and interrelated, factors: lack of information and the passage of time. With more time available, one can garner more information, for example from repeated experiments, and so risks would decline. With more information available, the future should become less unpredictable. However, what we want to concentrate on today is some aspects of the inter-temporal element of risk, rather than that which arises from lack of information. As Andrew Crockett, the General Manager of the BIS, stated in a speech at Basel last year (21 September 2000) (and previously in his keynote speech at the last Colloquium in Vienna, April 2000), that

> we think of risk evolving through time. The received wisdom is that risk increases in recessions and falls in booms. In contrast, it may be more helpful to think of risk as *increasing* during upswings, as financial imbalances build up, and *materialising* in recessions. The length of the horizon here is crucial.[1]

Even if outcomes did follow policy decisions almost immediately, risks would, of course, still occur, arising for example from lack of knowledge, failure to appreciate the context, principal/agent problems, and perhaps from instantaneously occurring shocks of one kind or another. But it should be much easier to hold decision-makers accountable for their decisions, because the outcome would become apparent while they were, by definition, still on the watch (and before conditions could, perhaps, change that much – though the world can change in a flash, as it did on the morning of Tuesday, 11 September).

Even so, any particular good outcome could not only be due to luck, but to an interaction between luck and a high level of risk-tolerance by the individual decision-maker. In this case the problem is that the principal, say a loan officer's boss or the bank's shareholder, does not have enough information on the agent's (i.e. the loan officer, the bank) characteristics, e.g. effort, ability and risk aversion. If each period's activity reaches its final outcome within that same period, and each period's game is independent from that of the previous and next game, then information on such characteristics will build up over time. This suggests,

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

for example, that incentives should be designed to reward average success over a long period of time rather than short-run extraordinarily good results. This would not suit those who do not expect or plan to be staying put in a single job or with a single firm for the long haul, but are these short-term characters the kind of agents that firms should want to hire?

In this respect we must confess to some puzzlement why the conventional wisdom regards it as acceptable for regulators to intrude upon, and to constrain, decisions relating to commercial banks' capital structure, but to abstain from any overt, or generalised, controls upon the design of such banks' bonus and payments' systems. After all it is the prospect of direct personal pecuniary reward that largely drives our decisions to expend effort and take risks, not the level of (operational) capital prescribed by the Basel Committee.[2]

As a generality, however, it appears to be remarkably difficult to make agents accept deferment of payment until the results of their actions become clearer. One can understand why. It can be difficult for a principal to commit credibly to a deferred payment, especially when the agent may have left the firm or even have died. Moreover the time lag before outcomes are finally revealed is uncertain, and some large part of such results will have been due to chance or to the decisions of others whom the agent cannot affect. Be that as it may, the convention is for the rewards of agents to be based on this period's outcome, even if that outcome is largely the result of decisions taken well beforehand.

Such outcomes in any period will also, of course, be due to the general business conjuncture. Profits rise and bad debts fall when the economy is strong, and when asset prices have been rising. At such times capital is easier to raise. So, typically, capital controls do *not* bite during booms. By the same token, during periods of weak growth and falling asset prices, profits fall and bad debts rise; capital is harder, sometimes impossible, to raise. Capital requirements are, by their nature, procyclical, and to some large extent, as exemplified by the last set of Basel Committee proposals, the more sophisticated is the measurement of individual firm risk the greater the consequential procyclicality in the aggregate.

While it has been generally accepted that additional sophistication in risk assessment at the individual bank level will add to greater aggregate procyclicality (see for example Kupiec, 2001 and Reisen, 2001), it is difficult to provide rigorous quantification, particularly when the necessary large-size databases are not available to the academic community. One on-going exercise, being undertaken at the BIS and using Mexican bank data, does confirm that the Basel proposals will lead to an increase both in required capital levels and in the procyclical variations of such requirements in that country (Lowe and Segoviano, 2002).

There is inevitably a conflict between the micro-level and the macro concerns in the operation of financial regulation (see Crockett, 2000 and Borio, Furfine and Lowe, 2001). At the micro-level there is no doubt that commercial banks and other financial intermediaries will tend to be more fragile, closer to insolvency, during recessions. Yet the real concern, the major externality, in banking lies in the possibility of contagious (and/or correlated) failure for a significant part of the system as a whole.[3] If a large number of banks are constrained by regulatory pressures to reduce their loans during a recession either to all (private sector) borrowers or to a significant subset (which may now be reclassified as higher risk), then the prospect of contagious collapse may even be enhanced (see for example Danielsson *et al*., 2001; Danielsson and Zigrand, 2001; Danielsson, Shin and Zigrand, 2002).

Is it possible for the regulators, and indeed for the firms themselves, to aim off, to adjust, somehow for the stage in the cycle in which they currently happen to be? Could some kind of cyclically adjusted capital requirement, or provisioning for bad debts, for example, be contemplated? In our view this may be extremely difficult. What one can tell is whether the rate of change of output, or of asset prices, has been greater, or less, than recent trends; but one can never be sure whether such trends may themselves be changing, as they do from time to time, as evidenced in productivity, equity prices, output, etc. For example, we have taken a diagram from a speech made by Laurence Meyer of the F.R. Board (Meyer, 2001) indicating that there were quite long periods of productivity growth both above

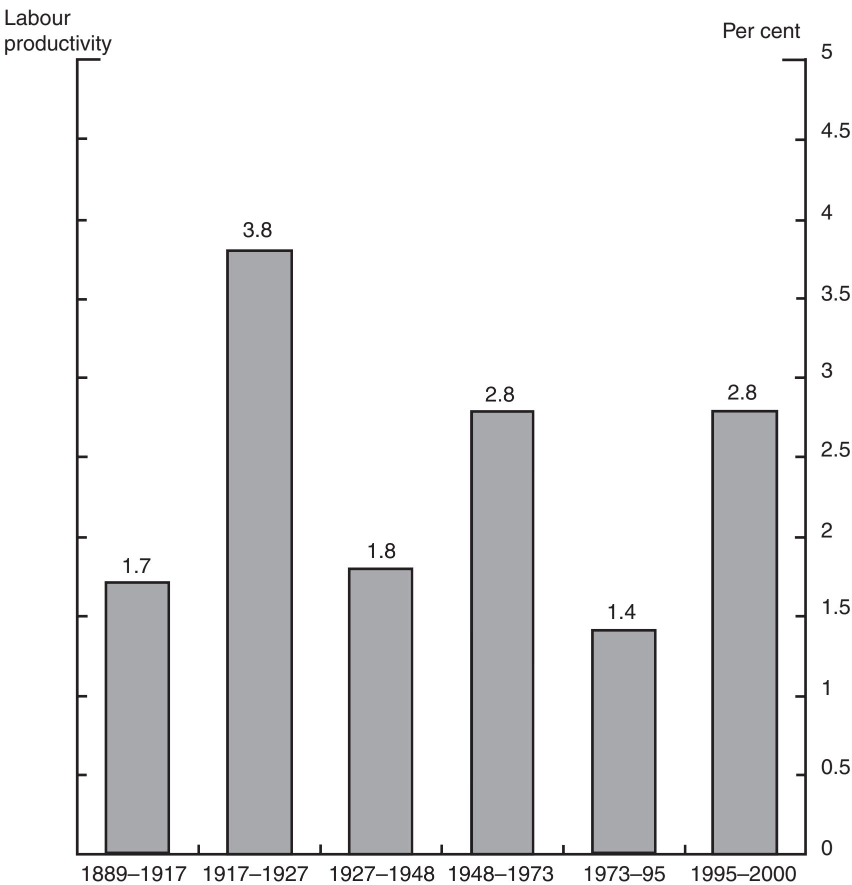


Figure 2.1 Productivity growth: the long view.
Source: Meyer 2001.

and below the long-run average of two per cent (Figure 2.1). He stated that to his eye, 'the charts suggest a sequence of waves in labor productivity, periods of rapid growth followed by periods of more sluggish growth', when 'the high-productivity periods reflect the influence of a bunching of technological innovations'.

Similarly one can decompose overall growth into periods of faster, and slower, growth. Take the UK, for example, though we have done similar exercises for the USA, Germany and Japan (not shown here). The most obvious interruption of growth came between the two World Wars, but even outside that interval, one can observe longer periods of faster and slower growth (Figures 2.2 and 2.3).

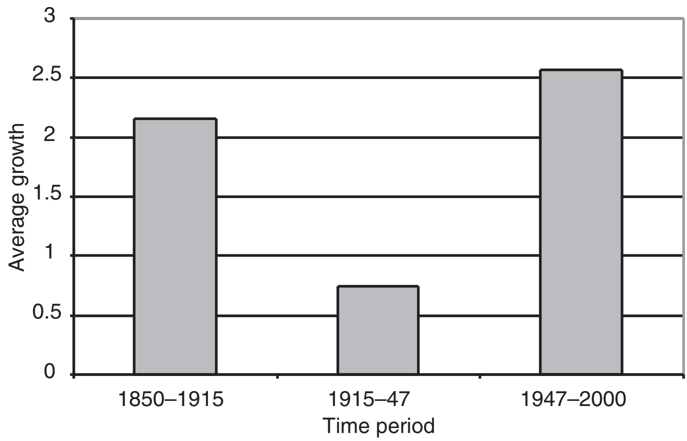


Figure 2.2 Average UK GDP growth (1850–2000).

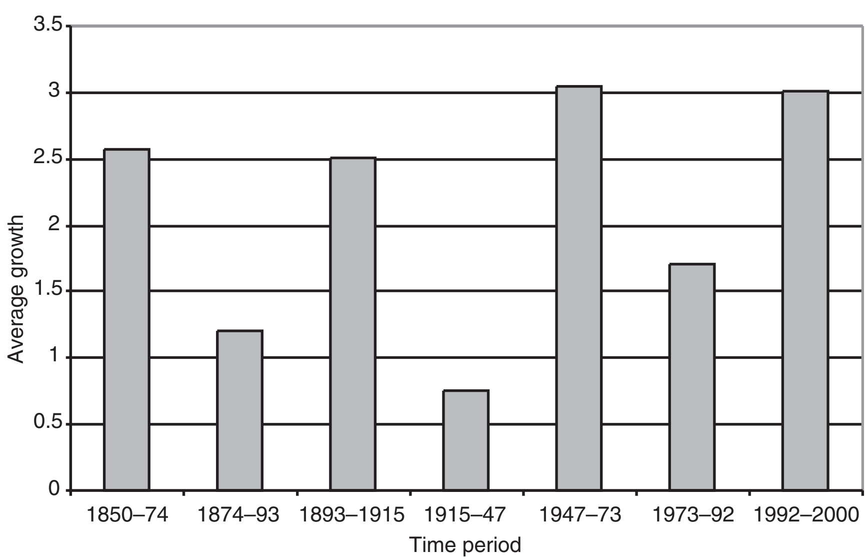


Figure 2.3 Average UK GDP growth (1850–2000).

Moreover, there have been much more marked longer-term variations in the growth rates of equity prices in real terms, i.e. deflated by the CPI. See, for example, data for the UK and USA (Figures 2.4 and 2.5).

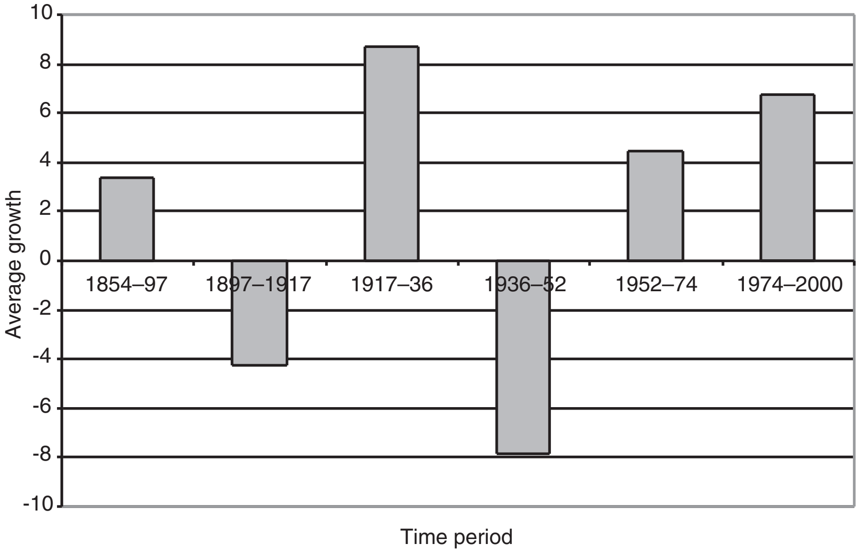


Figure 2.4 Average growth of UK real share prices (1854–2000).

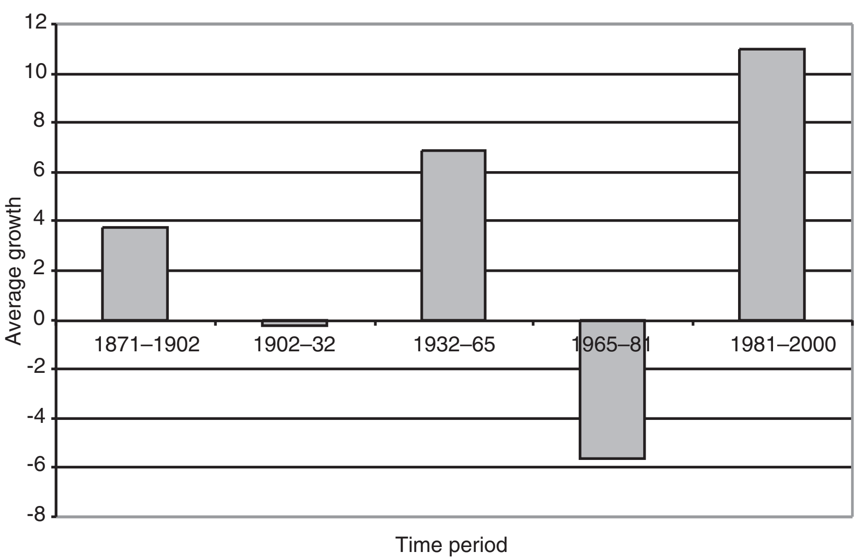


Figure 2.5 Average growth of US real share prices (1871–2000).

From that, it appears that the equity premium puzzle, the excess return to holding shares, is primarily a post-war phenomenon, a feature of the last two or three decades, and could, perhaps, now disappear as fast as it arose.

Even when the trends may be assumed, or taken, to be constant, so that fluctuations around them are stationary, with reversion to the mean in deviations around that trend, there is no good way to predict when the reversal will take place. Cycles do not have a constant periodicity, if indeed one can talk about cycles at all, rather than intermittent shocks which tend to have prolonged effects as a result of subsequent interactions within the economy. If one takes the available historical data, quarterly time series since about the 1950s, or annual data going back into the last century, there is very little sign of any constant regularity, or periodicity, in the onset of recessions. We have tried to formalize this assertion in two ways. First, if cycles have a regular periodicity, we should be able to reject the claim that the probability of a new recession aborting the recovery/boom is identical in each period following the initial onset of that recovery, irrespective of the preceding length of that recovery. If the occurrence of a new recession is independent of the length of the preceding recovery, i.e. the event is memoryless, then the number of such events in any fixed-length interval ought to follow a Poisson distribution, whereas if there was any cyclical regularity we should be able to reject the null of there being such a distribution. In our tests, using quarterly data with a fixed five-year window for the UK and Germany and a ten-year window for the US using data from the NBER ‘Business Cycle Dates’ database, we could not reject the null hypothesis that downturns did follow such a Poisson distribution, for any of these countries. This work is set out below in more detail in the Appendix, Part I, for which we thank Ryan Love.

Second, if the *timing* of a recession is independent of the length of the previous boom, then the time interval between the onset of the initial recovery and the subsequent next occasion of recession should follow an exponential distribution. Again we tested whether such time intervals did follow an exponential distribution, and over our data period could not reject that null hypothesis for the UK and Germany. We could, however, reject it for the USA, which does suggest that there the timing of recessions is not entirely independent of the length of the prior boom, which in our view is a pre-condition for belief in any cyclical regularity. Again this is set out below in more detail in the Appendix, Part II, by Love.

So the bulk of our historical evidence leads us to believe that one cannot predict the length of any upturn, or the likely future dating of the onset of a recession. One simply does not know where one might be in a cycle of indeterminate length.

Moreover, the fact that, say, output or real equity or housing prices have been rising faster than their previous trend will be taken by some as a consequence of a change in the fundamentals leading to a new higher trend, and by others as a sign of disequilibrium, or of misalignment. Even those who recognize a developing misalignment will have little, or no, idea when the so-called bubble may burst, or a bottom be formed after a burst. Economists, who by training try to find order in a complex system, are prone to see signs of misalignment early on; Alan Greenspan talked about ‘irrational exuberance’ as early as December 1996. If such misalignments last longer than we had expected, then we tend to wonder whether there has been a real change in the trend, a ‘new economy’. (Here I speak from experience. I took out a put option on the NYSE every year from 1995 to 1998, but then, baffled, gave up doing so.) The bottom line is that it is very difficult to condition regulation on the cycle, since we do not know where exactly in the cycle we actually may be at any moment.

Much the same is true when we turn from the first moment of asset prices, rates of return, to the second moment, volatility. What we know is that there is auto-correlation in volatility; when the market is calm it tends to be succeeded by quiet days, and vice versa. That is to say that a shock sets off a series of aftershocks. The well-known GARCH model fits the data for asset markets well. Although the auto-correlation is usually high, so that an initial shock often decays slowly, and in a few studies you cannot reject a unit root in volatility, an I-GARCH condition, the general finding is one of a gradual, long-drawn-out reversion to a mean.

But what we cannot now predict are the initial major shocks that can drive us from a basically quiet market to a very disturbed one. Whereas one can predict the aftershocks reasonably accurately, one cannot predict the initial shock, or hardly at all. We have examined daily data for the UK FT 30 index from the beginning of 1977 until end September 2001. Figure 2.6 shows the overall distribution of daily returns, and *overall* the time pattern of returns can be well modelled by a GARCH distribution indicating a high level of auto-correlation in volatility, clustering and time dependence. If, however, you isolate the extreme events, in this case with a probability threshold of 0.2 per cent or 1 in 500, i.e. roughly once every two years,

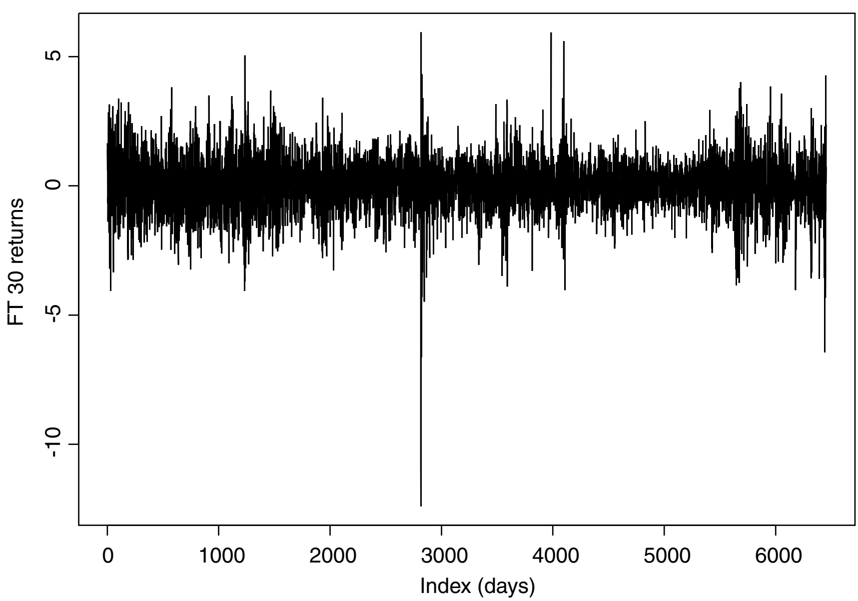


Figure 2.6 Daily 1977–2001 FT 30.

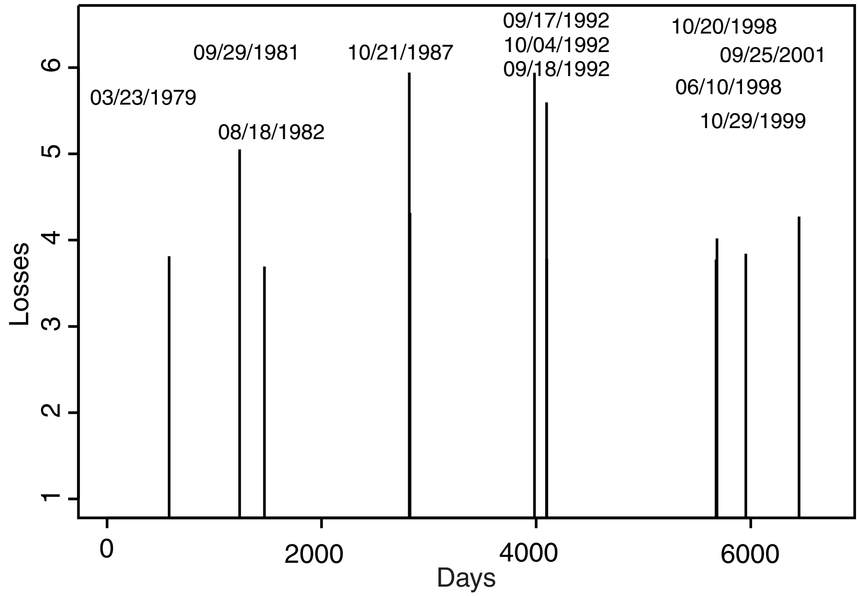


Figure 2.7 Daily 1977–2001 FT 30 extremes $p = 1/500$; p is probability.

the onset of extreme events is much more nearly random (see Figure 2.7). Even here, when you get a really massive shock as in October 1987, September 1992 or in early autumn 1998, the initial unpredictable shock sets off reasonably predictable aftershocks. Note also two features. First, as we move from the really extreme shocks back towards the middle of the distribution, so clustering, and time dependence, increase (see Figure 2.8). Second, in the same way as data on returns, growth, productivity, etc. do show long-term fluctuations, so there do appear to be long-run cycles in average volatility, and such volatility appears to be unusually high in equity markets currently, so that data on outliers from the distribution of volatility should really be conditioned on changes to its average level (Figure 2.9).

As with earthquakes, the aftershocks in markets that GARCH-type models do pick up so well will tend to bring down financial institutions and companies, especially those that have been weakened by the initial shock. So the danger for financial institutions is significantly higher in disturbed periods following some initial shock. But the danger of some *initial* major shock, whether it comes from outside the economic system as in August 1914 or on 11 September 2001, or from inside as in August/September 1998 or October 1987, is always lurking in the background. Consequently models that place their empirical basis on recent history, say the last year or so of market data, tend either to exaggerate risk, if one of these initial earthquake events has recently hit, or alternatively to be unduly complacent if markets have had quite a calm spell in the latest period.

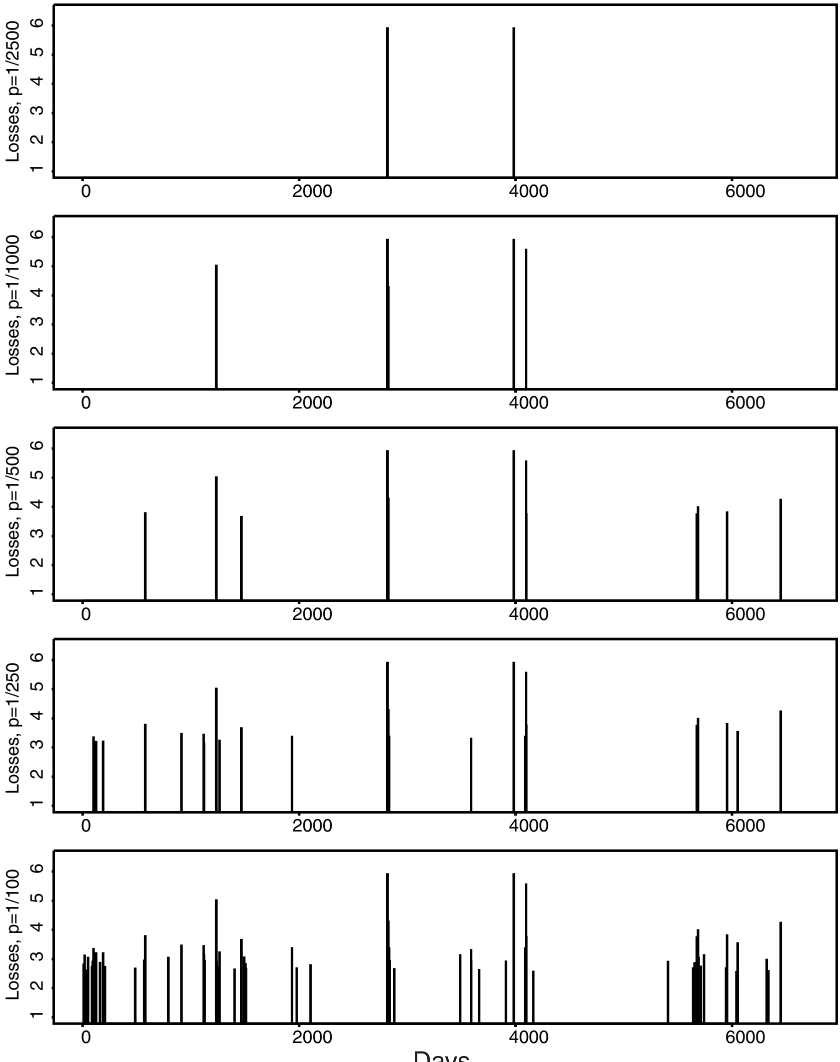


Figure 2.8 Moving from extreme shocks back towards the middle of the distribution.

Let us illustrate this with another diagram (Figure 2.10) taken from work by Jon Danielsson, in this case based on returns from the Hang Seng Stock Exchange Index in the final few months of 1997, which period included an extreme event, the speculative attack on the Hong Kong dollar and the authorities' response to that in raising interest rates extremely sharply. This shows the resultant Value at Risk (VaR) for a representative HK stock portfolio based first on a GARCH-type model which assumes that once a major shock has occurred that shocks will continue with only slow dampening, and second on a long-run, 500-day historical

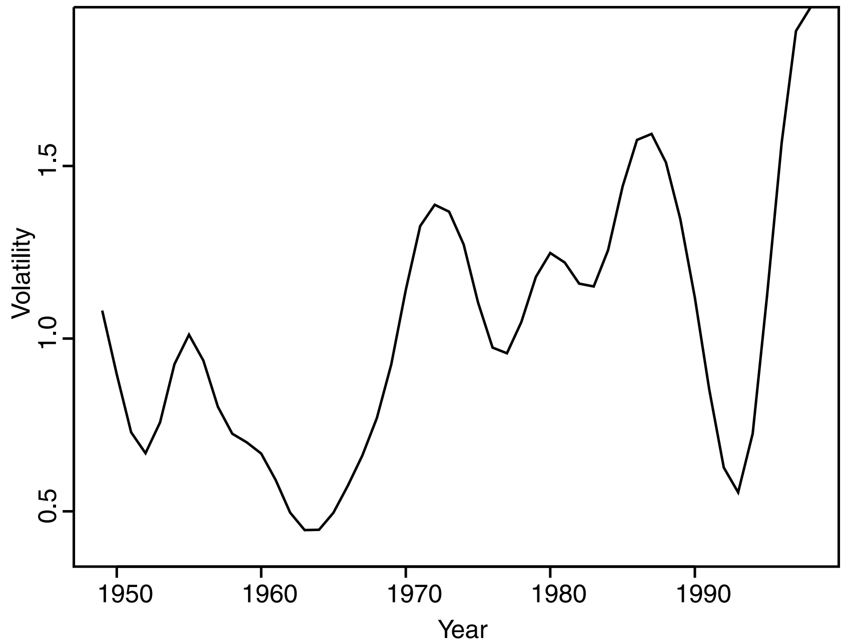


Figure 2.9 Long-run cycles in average daily volatility.

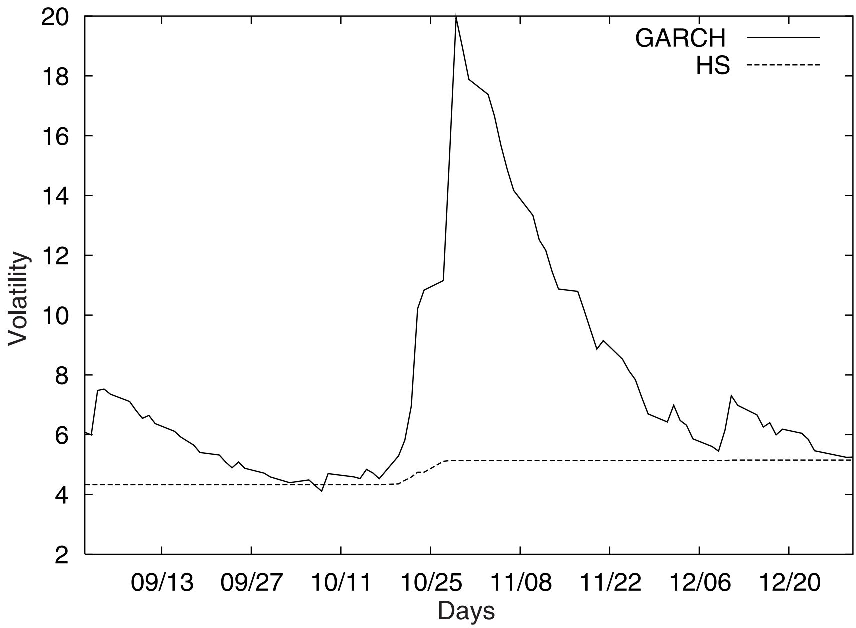


Figure 2.10 VaR on a 100mn. Portfolio of the Hang Seng Index, last quarter 1997, measured by GARCH and Historical Simulation (HS).

simulation, which weights the likelihood of extreme shocks in relationship to their past actual number of occurrences. The more that weight is placed on the immediate short-run, emphasizing the predictability of aftershocks, *not* the unpredictability of the major shocks themselves, the more is true risk likely to be overestimated during disturbed periods, especially in downturns, and equivalently underestimated in calmer periods.

Let us again emphasize that these extreme shocks, especially the downward shocks, are not predicted by markets. And there is a good reason for this. If we had appreciated that the Scandinavian housing bubble, the Japanese asset bubble, or the NASDAQ, would burst, then the earlier bubble would never have formed. Most disasters, economic and military, arise in some sense out of an intelligence failure. It follows from that that the internal market measures of risk, whether taken from option prices or deduced from other market prices, say bond spreads, can only help to predict the secondary aftershocks, not the main crises. Since such risk measures, like VaR and credit risk models, tend to extrapolate the recent past into the future, they too will lead to an unduly procyclical exaggeration of risk. If incorporated into regulatory constraints, the Calomiris-type proposals to constrain bank activity, for example, when bank bond spreads (relative to spreads on government debt) rose, could, unless radically revised, lead to financial paralysis whenever some outside major shock led to a general weakening of confidence. For example, we dread to think what would have happened in 1982 when the LDC debt crisis reverberated around the world, or in 1998 when the Asian crisis interacted with the Russian default and the LTCM collapse, if a Calomiris-type proposal had been in force and strictly followed. Note that after 11 September 2001, risk spreads everywhere were rising; but regulators were responding, and correctly so, by discretionary *easing* of regulations on insurance companies and on equity buy-backs by firms.

One of the most common human characteristics is to lock the stable door when the horse has already bolted. Those of you who have flown here will know what we mean. In most circumstances that propensity is relatively harmless, as well as entirely natural. But in the monetary field, the prospect of financial regulation doing the same could result in disaster by further restricting credit and liquidity in the aftermath of (often largely unforeseeable, and certainly unforeseen) crises.

Let us now try to gather up some of the threads. First, it is difficult to condition regulation on the current deviation of the economy, or key sectors of it, from the ‘fundamental equilibrium’, since we only get to know what that actually was after the event, and usually many years after the event. Second, we cannot hope to predict, or model, the really big adverse shocks, since these are almost by definition unpredictable. All we can do is model the aftershocks. Now this latter leads to a real dilemma. In the event of a serious adverse crisis, financial intermediaries *are* individually more fragile, but in aggregate you want them to be more expansionary. By the same token during an expansionary boom, individual banks are stronger, but in aggregate you would wish them to be more cautious.

How should, or can, regulators respond? One very common response to this, much used by the regulatory authorities, is that we have two objectives, individual firm solvency and systemic stability, and we have two main instruments, regulatory controls, primarily on bank capital, and short-term interest rates. In short, it is argued that it must be the responsibility of macro-monetary policy to vary interest rates sufficiently to maintain aggregate systemic stability, while it is the proper job of the regulator to adjust regulatory controls to the current actual condition of *individual* financial intermediaries, especially banks; and if that imparts a significant procyclical impulse, then so be it. It is the job of the Monetary Policy Committee (MPC) to offset that.

We understand that argument, and we appreciate that some find it powerful. But we do not share it for several reasons. Let us take first a more philosophical reason. We have many friends, colleagues and critics, especially among neo-classical liberals, who challenge the need for *any* regulation of banks and other financial institutions beyond those available to the consumer of any professional service. Our own answer to this has always been that banking for a variety of reasons is almost uniquely prone to an externality, in the shape of a contagious or highly correlated crisis. If it then really is to be the case that our banking regulations may well actually worsen the procyclicality of the system, make it more, not less, prone to systemic crises, what then is the point of this huge regulatory edifice, with its ever-increasing complexity and use of resources?

Our second point is more practical. This is that we do not believe that Monetary Policy Committees can be expected to vary interest rates, against the grain of a procyclical regulatory system, sufficiently to maintain systemic stability. There are several strands to this argument. First, central banks are committed to inflation, or price targets, with such targets defined in terms of current goods and services. Many of the problems of financial stability arise from fluctuations in asset prices. Although asset price fluctuations do affect goods and services inflation, they only do so to a limited extent, so that the response of MPCs to asset price movements, by themselves, is usually somewhat hesitant. Second, nobody ever can convincingly state whether, and how much, current asset prices are misaligned. So even if a central bank had a mandate to respond when markets showed undue exuberance, there would be a great hesitation to do so. If you cannot condition regulation on asset market misalignments, it is almost equally difficult to condition macro-monetary policy on them. Third, there are strong arguments made, with some force, that regulatory constraints ought to be subject to pre-commitment, to avoid forbearance, etc. On the other hand, macro-monetary policy is discretionary. If regulators were, quasi-automatically, to close down (some of) the functions of (systemically important) banks, the resulting shock to confidence could overwhelm the limited, and possibly delayed, counter-measures of interest rate changes. Fourth, members of MPCs usually tend to be cautious folk; the idea that regulation can be procyclical, or become even more so, because we will all be rescued by the rapid, expert and aggressive variations in interest rates ordained by the brilliant men and women members of our MPCs, is not one that we personally find enormously appealing.

If we reject the argument that we can safely leave systemic stability to the good offices of the world's MPCs, what then do we do? This, of course, is the point at which we hand you back to our Chairman. But we do have a few ideas, which we hope that you will not find too scatty.

First, our view is that markets are much better at setting *relative* prices, than being able to fix aggregate fundamental equilibria, and much better at assessing relative volatility (and risk) than aggregate values. This is because, with respect to returns, the major shocks, in both directions, are basically unpredictable; and also because, with respect to volatility, the interaction of the natural progression through a sequence of jobs and roles during the ageing process means that those taking trading decisions usually have a relatively short historical experience. That, combined with a natural myopia, means that recent short-run experience is likely to be given excessive weight in both asset pricing and risk assessment. That, in a nutshell, is why we would tend to be extremely wary of the proposals for Full Fair Value Accounting (FFVA) (see also Saurina, 2001 and Commission Bancaire, 2001).

What that implies, in our view, is that regulators should place much less weight on the means of distributions, e.g. of asset values, or of spreads over riskless assets, or of the various VaR or credit models, especially when based on short runs of data, and much more attention to the relative position of the various banks and financial intermediaries within the distribution of such models, whether overall those measures seem strong or weak.[4] In so far as there are market signals that can be observed, the regulators should, perhaps, be continuously focusing on those in the worse tail of the distribution, say the bottom twenty per cent, whether the current state of the barometer, and the absolute readings, indicate that the current weather conditions are fair or stormy. Those conditions will change, and at times and in ways that cannot be predicted. Of course, there may be occasions when a large proportion, perhaps almost all, of a country's banking system will need close attention, as for example in Japan, but such occasions are usually obvious.

The second point is that the interaction of occasional large unpredictable shocks (interspersed by GARCH-type predictable aftershocks), with some natural tendency to myopia, means that regulators need to make strenuous efforts to lengthen the horizon over which regulating metrics and decisions are made. The various kind of risk models, VaR and credit risk measures, bad debt provisioning, etc. need to be based on the longest runs of data available. This is not so much, to use the catchphrase, 'to look through the cycle', since on this view the concept of any regular cycle in asset prices or in the economy is certainly overstated, and quite probably wrong-headed in the first place. It is rather to ensure that the metrics are constructed over a sufficient length of time to incorporate both bad/volatile periods including some stochastic bunching of severe adverse shocks and also good/calmer periods.

What the regulator has to contend with is the tyranny of the annual cycle of accounts, salary and bonus negotiations, and tax payments. The employee, the accountant, the shareholder and the tax official all demand payment flows conditioned on actual, ex post annual data. But the relevant periodicity for the regulator should not be shoehorned into such arbitrary and short slices of time. Is it, for example, possible to build on the Spanish experience with preprovisioning, and make various aspects of regulatory requirements a function of much longer time-average experience (also see Borio and Lowe, 2001a; Borio, Furfine and Lowe, 2001; and Commission Bancaire, 2001); and, to return to an earlier theme, insist on some longer-run time averaging of incentive and bonus schemes as a key feature of the overall regulatory framework?

A final consideration is whether one can condition regulatory requirements on the first differences of economic outcomes, e.g. output and (real) asset prices, rather than on levels. For the reasons already outlined we just do not believe it to be possible for regulators, or for anyone else, to determine with any certainty where the economy might be in the so-called cycle, or to determine quantitatively the extent of asset price misalignment. But we can measure, by simple arithmetic, past trends, and we can estimate with reasonable accuracy whether current growth is above, or below, those trends. So, collateral requirements, loan-to-value ratios, minimum capital requirements, etc. could all be raised when such increases, e.g. in asset prices, were above trend, and lowered when there was a recession relative to trend. Of course, trends can change, but the continued recalculation of long-term trends, and the falling out of the calculation of earlier first differences as time passed, would reduce the danger of resultant error.

For example, how might a system of required minimum loan-to-value ratios look like that were conditioned on fluctuations in housing price growth around its trend; or capital adequacy requirements that were conditioned on fluctuations in bank credit expansion (to the private sector) around its trend? Note that there is ample evidence, from many countries and both at the micro and macro level, that such earlier over-rapid credit expansion is a potent harbinger of current fragility (for two recent studies on this, see Niinimäki, 2001 and De Lis, Pagés

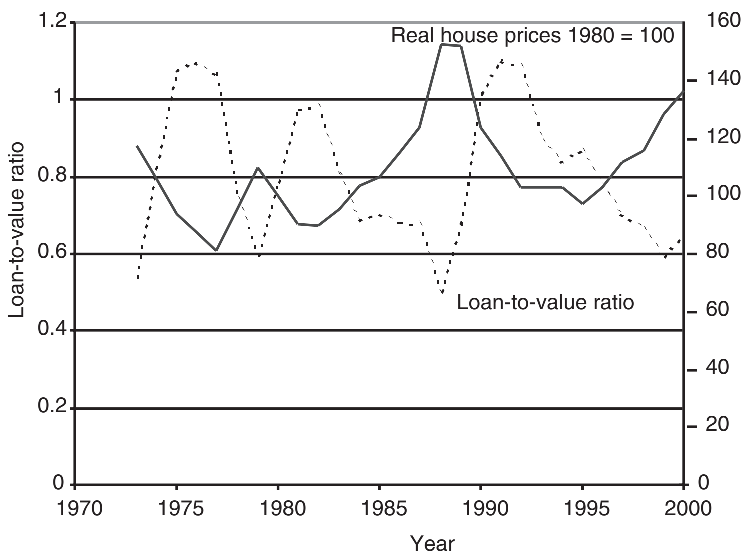


Figure 2.11 Real UK house prices and estimated loan-to-value ratios.

and Saurina, 2000). Anyhow, it is relatively simple to construct a simulated system whereby such conditioning results in apparently sensible counter-cyclical movements in the regulatory variables (see Figures 2.11 and 2.12). A more complete description of the simulations run is set out below in the Appendix (by Ryan Love).

To conclude, we are not confident that any of these suggestions would stand much chance of practical acceptance, or, if accepted, would be sufficient to offset the procyclical nature of our prospective regulatory system, but we do need to consider these issues seriously.

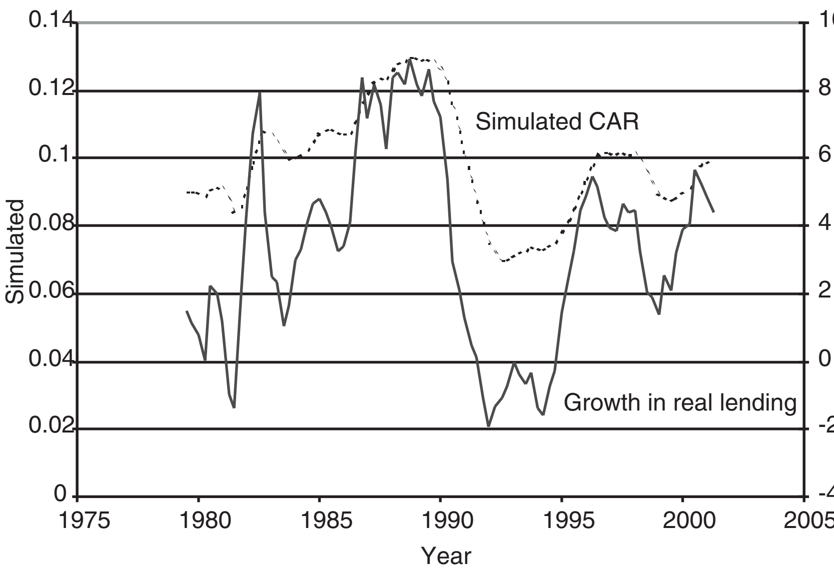


Figure 2.12 Growth in real lending and simulated capital adequacy ratios (CAR).

Appendix 2A (by Ryan Love)

Part I: Randomness of economic downturns

What we are seeking to determine is whether the event of an economic downturn is independent of what has happened previously. The hypothesis to be tested is that the probability of there being a downturn in economic activity in the next quarter, using quarterly GDP data, is independent of whether there was a downturn last quarter or ten quarters ago. Due to the nature of the data, there are two linked methods of testing this: the use of the Poisson and exponential distributions. Both distributions have the important property of being memoryless. For example when we examine the number of occurrences, in this case downturns, in fixed-length intervals in the period under investigation, the Poisson distribution will hold if the number of occurrences in (a statistically sufficient number of) non-overlapping time intervals are independent of one another. The other conditions that must also hold in order for the Poisson distribution to be followed are:

1 The events (downturns) occur at random in continuous time. Although our results use quarterly data, our tests, explained in more detail below, use five- or ten-year windows. Therefore with twenty or forty quarters in each window, this is 'large' enough for our tests to be valid.
2 The probability of two events occurring simultaneously is zero. This is clearly the case as one economic downturn cannot occur at the same time as another.
3 Events occur uniformly so the expected number of downturns in a 50-year period is twice that expected in a 25-year period. Again this appears reasonable.

For this discrete distribution, the probability distribution is given by:

$$P(x) = \frac{e^{-\lambda} . \lambda^x}{x!} \qquad x = 0, 1, 2, \ldots$$

$$\text{mean} = \text{variance} = \lambda$$

x, the random variable, is the number of downturns in any fixed interval of time (set arbitrarily). The justification for using the Poisson distribution is based on the idea of 'the law of rare events'. If we have a number of Bernoulli trials when there is a small probability of an economic downturn then the familiar result will hold that the number of events (downturns) occurring follows a Poisson law.

The exponential distribution, instead of modelling the number of occurrences within a given time interval, considers the time between events in the entire period. Indeed if the number of events follows a Poisson distribution then the time intervals will follow an exponential distribution. The theory is commonly used in queuing phenomena, for example the time between machine breakdowns or waiting times at hospitals. The theory can be used in the same way in the sense that we are 'waiting for a recession'. In all cases of exponential processes, the waiting time until one event is independent of the waiting time until the previous event, which is indeed what we are trying to test. The PDF and CDF for the exponential distribution are given by:

$$\text{PDF} \quad f(x) = \frac{e^{-(x/\mu)}}{\mu} \qquad x \geq 0$$

$$\text{CDF} \quad F(x) = 1 - e^{-(x/\mu)}$$

$$\text{mean} = \mu, \text{ variance} = \mu^2$$

The similarity between the Poisson and exponential distributions is shown by the relationship between the parameters, $\mu = 1/\lambda$.

Therefore, in order to test the hypothesis that economic downturns are independent of one another formally, we can fit the data to both the Poisson and exponential distributions and test whether the data appears to come from these processes. If we cannot reject the hypothesis that the number of downturns in any given interval comes from a Poisson process, or that the length of the time between downturns is generated by an exponential distribution, then we can argue that the probability of there being a downturn in economic activity is independent of what happened previously. In other words the event of a recession or downturn is indeed independent of what has gone before – it is memoryless.

Unfortunately, since the number of recessions in the data is very small,[5] there are insufficient data to test whether *recessions* are memoryless. However to get around this problem we can test whether *downturns*, defined as one (as opposed to two) or more quarters of negative GDP growth, are memoryless. Although the number of single quarter downturns are much more numerous, they are still not as frequent as we would ideally like when examining each country individually. This may result in the power of the tests not being particularly high, i.e. not being able to reject the null hypothesis when it is in fact false. For this reason we not only consider the UK, the USA and Germany separately, but also combine the data from all three countries.

GDP growth rates were calculated for each quarter and were standardized by various multiples of the growth standard deviation. For the UK and Germany the data were broken down into five-year intervals and the number of downturns greater than x standard deviations was observed. Too large a value of x would result in too few significant downturns to test any hypothesis, whilst too small a value of x would result in periods of negative growth being considered in the test that are not significant enough to be classed as true economic downturns. Therefore different values of x were used, ranging from 0.25 to 2. The frequency of 0, 1, 2, …-downturns in these five-year intervals was then tabulated and the frequency of downturns if the data followed a Poisson process was also generated. For the USA, the data were taken from the NBER database, which gives the dates of USA economic downturns from 1854 to the present day. With no quarterly GDP figures available over the earlier years, no standardization could be undertaken and so the number of downturns, as defined by the NBER, in ten-year intervals from 1860 was used. The test statistic to be used is given as:

$$K = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i is actual frequency, E_i is expected frequency should the data come from a Poisson process and n is the number of categories (0 downturns, 1 downturn, …, $n-1$ downturns). Under the null hypothesis that the data are consistent with a Poisson process, K should be distributed as a χ^2 with degrees of freedom equal to n minus [1 plus the number of parameters to be estimated, in this case, 1 for λ], i.e. $n-2$. The results are given in Table 2A.1.

As can be seen, the null hypothesis that the data are generated from a Poisson process, i.e. that an economic downturn is independent of what happened previously, cannot be rejected even at the ten per cent level for all three countries separately and also when combining the data. From this we can suggest that the event of an economic downturn is independent of previous economic activity, i.e. economic downturns are memoryless. Such conclusions can, however, be criticized on the basis of our small dataset.

Table 2A.1 Testing for the randomness of economic downturns: Poisson test

	Standardized by x standard deviations: x =	*Number of significant downturns*	*K*	*Degrees of freedom*	*Critical values 10%*
UK	0.5	23	4.40	4	7.78
1955–2000	0.75	13	1.38	3	6.25
	1	10	0.05	2	4.61
	1.5	5	0.01	1	2.71
	2	3	1.55	1	2.71
USA 1860–2000		30	3.26	3	6.25
Germany	0.25	26	2.83	4	7.78
1960–2000	0.5	13	1.62	2	4.61
	0.75	10	1.68	2	4.61
	1	5	0.06	1	2.71
All	0.5	44	5.96	3	6.25
1960–2000	0.75	30	4.05	2	4.61
	1	21	2.02	1	2.71

Note: Five-year intervals were used for the UK and Germany, ten-year intervals for the USA and two-year intervals when considering all three combined.

When testing the data against the exponential distribution, the time between economic downturns was calculated (in quarters) and the frequency of one quarter, two quarters, … intervals between downturns was computed. The expected frequency was also found, assuming the data were generated from an exponential distribution, and the K statistic was computed in the same manner as before.
The results are given in Table 2A.2:

Table 2A.2 Testing for the randomness of economic downturns: exponential test

			Critical values		
	K	*Degrees of freedom*	*10%*	*5%*	*1%*
UK	3.38	7	12.02		
USA	18.54	6			16.81
Germany	10.88	6	10.64	12.59	
All	17.01	8		15.51	20.09

The null hypothesis that the time between economic downturns is consistent with the exponential distribution (the time until one downturn is independent of the time until any previous downturn) cannot be rejected for Germany at the five per cent level and even at the ten per cent level for the UK but the hypothesis *can* be rejected at the one per cent level for the USA and at the five per cent level for the three countries combined.

The results from our tests using the exponential distribution are not likely to be as robust as those when considering the Poisson distribution, not just because of the small dataset, but also because the exponential is a continuous time distribution and we are forced to take a discrete time approximation, i.e. integer values of times between downturns. Although this is a valid method, the results will not be as accurate as the Poisson results and so should be considered as a 'rough and ready' guide.

Part II: House prices

The following exercise examines whether growth in real house prices, i.e. nominal house prices deflated by the retail price index, can be found to be above or below past trends. First using annual UK data from 1953 we calculate a 20-year moving average of real house price growth rates, which we define as trend. For each year from 1973 we then compare the recent growth rates to this trend. This is done by summing the absolute deviations from trend for the current and previous two years' growth rates. However, in order to place greater importance on the most recent figures, each year was given a weight: 0.5, 0.3 and 0.2 for current year, one-year lag and two-year lag respectively, i.e. the following statistic was calculated:

$$S_t = 0.5\,(G_t - T_t) + 0.3(G_{t-1} - T_t) + 0.2\,(G_{t-2} - T_t) \qquad t = 1973 \ldots 2000$$

G_t is the growth of real house prices for year t and T_t is the 20-year moving average:

$$T_t = \frac{1}{20} \sum_{i=t-19}^{t} G_i$$

S_t then gives a measure of how real house prices are growing relative to trend. From this we then simulate a proxy for the loan-to-value ratio that would 'ideally' have occurred if lenders had behaved with due caution, based on past evidence. From broad-brush historical data, we assume that the average loan-to-value ratio was about 0.8, and again assume that reasonable lower (most conservative) and upper (most expansionary) bounds might be 0.5 and 1.1. Intuitively one would expect the loan-to-value ratio to fall when real house prices are rising as lenders should anticipate the inevitable end to recent above-trend house price increases, i.e. they expect growth rates to revert back to their 20-year trend. The loan-to-value ratio for year t was then approximated as:

$$LV_t = a + b(-S_t)$$

where a and b were calculated from:

$$0.5 = a + b(\min(-S_t))$$
$$1.1 = a + b(\max(-S_t))$$

$a = 0.785$ and $b = 0.062$. A graph showing real house prices and the estimated loan-to-value ratio is given in Figure 2.11, shown on page 31. As can be seen, and as one would expect, the ideal loan-to-value ratio would rise when real house prices fall and vice versa.

Part III: Bank lending

A similar exercise was carried out using quarterly data on bank lending to the private sector. Due to the notoriously imprecise estimates of trend/steady state level of economic activity when using GDP data,[6] bank lending figures were used, the idea being that bank lending will increase when economic activity, as estimated by consumer and investment spending, rises as individuals and firms borrow to finance their rise in expenditure.

Quarterly UK data were available from the fourth quarter of 1958 until the second quarter of 2001. Nominal figures for bank lending to the private sector were deflated by quarterly RPI to obtain real lending figures, and using the growth rates from the same quarter in the previous year, a 20-year moving average of growth in real lending was calculated and defined as trend. A measure of deviation from trend was then calculated by summing a weighted average of current and previous eleven quarters' deviations of growth from trend. This, defined as D_{12}, is given below:

$$D_{12,t} = \sum_{i=0}^{11} \alpha_i (G_{t-i} - T_t) \qquad t = 1979{:}4 \ldots 2001{:}2$$

Where G_t is the growth in real bank lending to the private sector and T_t is the eighty quarter moving average:

$$T_t = \frac{1}{80} \sum_{i=t-79}^{t} G_i$$

α_i are the weights associated with each quarterly deviation from trend and in order to place greater weight on the most recent quarter, a gentle quadratic decay of weights was used, satisfying the condition that the weights sum to unity. The weight placed on the current quarter was 0.22, that on four quarters ago, 0.10 and that on eight quarters ago, 0.03.

Again we have assumed that the preferred required capital adequacy ratio might be ten per cent, and that it might vary, under normal circumstances, between seven per cent (in a recession) and thirteen per cent (in a boom). Using the definition of deviations from trend, $D_{12,t}$ above, we perform a similar exercise, as done with the loan-to-value ratio, that gives the simulated movements in the required CAR as shown in Figure 2A.1 and Figure 2.12 on page 32. The simulated capital adequacy ratio was approximated as:

$$CAR_t = c + d\left(D_{12,t}\right)$$

where c and d were calculated from:

$$0.07 = c + d\left(\min(D_{12,t})\right)$$
$$0.13 = c + d\left(\max(D_{12,t})\right)$$
$$c = 0.093 \text{ and } d = 0.0069.$$

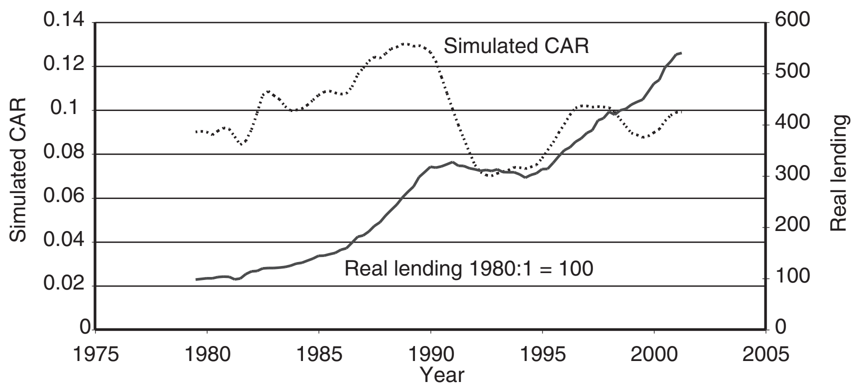


Figure 2A.1 Real lending and simulated capital adequacy ratios.

Notes

1 Borio and Lowe (2001b) report that 'One of the relatively few robust findings to emerge from the literature on leading indicators of banking crisis is that rapid domestic credit growth increases the likelihood of a problem'.

Similarly a study of bank failures during the small banks' crisis in the UK in 1991 and afterwards (Logan, 2001), found that 'rapid loan growth in the previous boom was found to be a good longer-term indicator of failure. Thus, unlike the survivors, banks that subsequently failed exhibited a pronounced boom and bust cycle in lending growth'.

(Also see Carmichael and Esho, 2001 and De Lis, Pagés and Saurina, 2000.)

2 Of course, pecuniary return is not our only driving force. The applause of our fellows and a grasp of power are two other key factors. For example, I advocated paying central bankers by result, when adviser to the RBNZ in the run-up to their Act in 1989, well before Carl Walsh (1995) formalized the idea. But neither he, nor I, have got very far in practice; usually we are told that the idea is somehow not seemly. Central bankers will give of their best, whatever the payment structure. No doubt, though, I would myself find it somewhat more reassuring if central bankers were paid by results, and paid handsomely when their targets were achieved.

In fact, I have been led to understand, the main reason why such an incentive scheme was not introduced in New Zealand was presentational; that is the NZ Treasury were concerned about possible headlines, after an interest rate increase, along the lines of 'Governor makes himself $20,000 by throwing 20,000 out of work'. In this case, unusually, the time lags may help in putting in place the correct incentive structure. After all, interest rate changes now only affect inflation in about two years' time, by when anyhow the immediate output effect will tend to be declining. So any bonus should only be earned after two years, by which time it will become somewhat clearer whether the initial decision was correct, or not.

3 At the BIS they think about the macro problem slightly differently. The main problem in their view is that banks have correlated portfolios and that if multiple institutions fail at the same time the macro or social costs outweigh the costs to the banks alone. Arguably, bank management do not take into account these additional macro costs in making decisions. Normally, the problem is not so much 'contagious' failures, but 'correlated' failures.

4 When I first started to work in the Bank of England, regulation, such as it was, tended to take the form of looking at what the better, or best-run, banks were doing, and then trying to get the rest to emulate them. We, keen young academics, tended to scoff at that, but perhaps there was more wisdom in the relative approach, than trying to adopt absolute standards in a shifting and unpredictable world.

5 A recession is defined as two or more consecutive quarters of negative GDP growth.

6 This imprecision arises partly because of the frequent revision of GDP data and also due to the ad hoc nature of trend estimation, either by Hodrick-Prescott filtering, taking first differences or using a linear trend.

References

Borio, C., Furfine, C. and Lowe, P. (2001) 'Procyclicality of the financial system and financial stability: issues and policy options', BIS Papers no. 1.

Borio, C. and Lowe, P. (2001a) 'To provision or not to provision?', *BIS Quarterly Review* (September).

Borio, C. and Lowe, P. (2001b) 'Asset prices, financial and monetary stability: exploring the nexus', paper prepared for World Bank–FRB Chicago Conference on Asset Price Bubbles, deferred from October 2001 until Spring 2002, draft dated 15 October.

Carmichael, J. and Esho, N. (2001) 'Asset price bubbles and prudential regulation', paper prepared for World Bank–FRB Chicago Conference on Asset Price Bubbles, deferred from October 2001 until Spring 2002, Australian Prudential Regulation Authority (September).

Commission Bancaire (2001) 'Dynamic provisioning and interest rate risk disclosure: an alternative to fair-value accounting for the banking book' (August).

Crockett, A. (2000) 'Marrying the micro- and macro-prudential dimensions of financial stability', paper presented at the International Conference of Banking Supervisors, Basel (21 September); also see his keynote speech at the previous SUERF Colloquium, Vienna, April 2000.

Danielsson, J. (2001) 'The emperor has no clothes: limits to risk modelling', Financial Markets Group, Special Paper no. 126, London School of Economics. Online http://www.riskresearch.org, forthcoming *Journal of Banking and Finance*.

Danielsson, J., Embrechts, P., Goodhart, C., Keating, C., Muennich, F., Renault, O. and Shin, H.S. (2001) 'An academic response to Basel II', Financial Markets Group,

Special Paper no. 130, London School of Economics (May).
Danielsson, J., Shin, H. and Zigrand, J.P. (2002) 'The impact of risk regulation on price dynamics', mimeo, London School of Economics. Online http://www.riskresearch.org
Danielsson, J. and Zigrand, J.P. (2001) 'What happens when you regulate risk? Evidence from a simple equilibrium model', mimeo, London School of Economics. Online http://www.riskresearch.org
De Lis, S.F., Pagés, J.M. and Saurina, J. (2000) 'Credit growth, problem loans and credit risk provisioning in Spain', Bank of Spain, Working Paper no. 18.
Kupiec, P. (2001) 'The New Basel Capital Accord: the devil is in the (calibration) details', International Monetary Fund.
Logan, A. (2001) 'The United Kingdom's small banks' crisis of the early 1990s: what were the leading indicators of failure?', Bank of England Working Paper no. 139 (July).
Lowe, P. and Segoviano, M. (2002) 'Internal ratings, the business cycle and capital requirements: Some evidence from an emerging market economy', BIS internal draft (February).
Meyer, L. (2001) 'What happened to the new economy?', speech to the New York Association for Business Economics and Downtown Economists, 6 June.
Niinimäki, J-P. (2001) 'Should new or rapidly growing banks have more equity?', Bank of Finland Discussion Paper no. 16 (September).
Reisen, H. (2001) 'Will Basel II contribute to convergence in international capital flows?', OECD Development Centre.
Saurina, J. (2001) 'Full fair value accounting implications for financial stability', Banco de España.
Walsh, C. (1995) 'Optimal contracts for central bankers', *American Economic Review* 85: 1.

3 Technology and finance: challenges for financial markets, business strategies and policy markets

New challenges for European and national regulations relating to financial markets*

Didier Reynders and Jean-Paul Servais

Policy makers face the challenge of the continuous development of the financial markets in terms of infrastructures, new products and alliances. This challenge is expressed not only in terms of the quality of the regulations relating to financial markets and services, but also in terms of a new definition of the hierarchy and the nature of the legal standards in this field.

Hyper-regulation as well as self-regulation have both proved their limits in this matter. This is therefore not an easy exercise, not only at the European level but also at the national level. It is at the same time a delicate and significant task in that the quality of the legal environment is destined to become more and more a competitive advantage for economic areas such as the European Union and the United States.

In this perspective, this paper first examines the main themes of the action of the Belgian presidency of the European Union in the field of financial services (taking these constraints into account), and second, describes some recent national developments.

The Belgian presidency was highlighted by a memorable event – the introduction of euro coins and notes. Such an operation involves many technical problems that needed to be resolved with care. Apart from these logistical aspects, the change of currency is of great symbolic significance. The Belgian presidency has used this unique opportunity to make the public more aware of all the advantages of European integration, to make progress in integrating economies and markets, and to identify potential ways of improving the functioning of all European institutions and their dialogue with the citizens of Europe. The future of Europe, where 300 million people share a common currency, depends on the coordination of national economic policies. It is essential to make a constant effort to ensure that the young people of Europe are aware of that fact.

The economic and monetary aspects of the Belgian presidency's achievements bear the stamp of Belgium's will to continue the implementation of the Financial Services Action Plan, leading to the creation of an integrated European market in financial services by no later than 2005. With regard to the strategy for

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

the execution of this plan, the Committee of Wise Men (chaired by Baron Lamfalussy), in its report that was approved by the Stockholm Summit on 23 and 24 March 2001, rightly pointed out that

> The European Union has no divine right to the benefits of an integrated financial market. It has to capture those benefits by building an integrated European market – in many areas starting from a very low level. If it does not succeed, economic growth, employment and prosperity will be lower, and competitive advantage will be lost to those outside the European Union. And the opportunity to complement and strengthen the role of the euro and to deepen European integration will be lost.

In order to put in place this proactive approach to building an integrated financial market, the Wise Men Committee stressed that

> change in the legislative process is of central importance. Not just to ensure the prompt delivery of the Financial Services Action Plan but so that in the future Europeans will be able to rely on a more accountable and efficient regulatory structure that will be able to match the best in the world.

All the above reasons have led to a reform (being implemented) of the decision-making process in the field of European Financial Regulation. This is commonly called the 'four-level approach' of the Lamfalussy Committee, i.e. a four-level structure of the financial law sources.

Level 1: Framework principles to be decided by normal EU legislative procedures (i.e. proposal by the Commission to the Council of Ministers/European Parliament for co-decision).

Level 2: Establishment of two new committees – a EU Securities Committee and a EU Securities Regulators Committee (the former FESCO) to assist the European Commission in determining how to implement the details of the Level 1 framework.

Level 3: Enhanced cooperation and networking among EU securities regulators to ensure consistent and equivalent transposition of Level 1 and Level 2 legislation (common implementing standards).

Level 4: Strengthened enforcement, notably with more vigorous action by the European Commission to enforce Community law, underpinned by enhanced cooperation between the Member States, their regulators, and the private sector.

This approach recognizes two layers in the legislation related to financial markets: basic political choices that can be translated into broad but sufficiently precise framework standards (Level 1); and the more detailed technical measures, in full conformity with this framework, needed to implement the objectives pursued by the legislation (Level 2).

Level 2 is composed of an actively functioning network of national securities regulators, the European Commission and a new European Securities Committee. It has to define, propose and decide on the implementing details of the so-called 'framework' directives and regulations determined by the co-decision procedure. This level enables the building of solid triangular relations of trust and efficiency between the European Commission, the Securities and the Regulators Committees.

Already in the first months of the Belgian presidency, the European Securities Committee (ESC) as well as the European Securities Markets Regulators Committee (ESMRC) have become fully operational. Indeed, both Committees met for the first time on 21 and 11 September respectively.

The Belgian presidency has paid careful attention to the need of exchanging views with representatives of the European Parliament on different draft directives and regulations in the field of financial services, as quickly as possible. Thus, as far as the projects belonging to the priorities of the Belgian presidency were concerned, all the concerned reporters of the European Parliament were contacted. Such contacts, even informal, made it possible to answer to a lot of questions of the European Parliament and, at the same time, to 'clarify' delicate aspects. This working method has the unquestionable advantage of nurturing a collaborative dialogue between the concerned European institutions, right from the start of the legislative process, and of avoiding in this way the multiplication of amendments. Thanks to this informal procedure, on 16 and 17 October, the Belgian presidency was able to reach an agreement on the 'directive on money laundering', a very sensitive issue, between the Council and the delegation of the European Parliament. This directive was finally adopted in record time.

As far as draft directives and regulations are concerned, the priorities of the Belgian presidency were the following:

- proposal for a directive on the prospectus to be published when securities are offered to the public or admitted to trading;
- proposal for a directive on the prudential supervision of financial conglomerates;
- proposal for a directive on inside trading and market manipulation.

These are all draft proposals in response to requests from numerous operators (most of whom engage in cross-border activity), but also from the various supervisory authorities, intermediaries and consumers, all concerned to see effective formulas covering these important aspects of the activity of businesses raising finance from the public.

- Draft regulation on the application of International Accounting Standards (IAS).

This draft (which aims to introduce a mandatory standard accounting system, at least for the compilation of the consolidated accounts of companies listed on the stock market) chimes perfectly with the idea that 'global players' should all have the same status in company law, in accounting law and in terms of the regulation of stock markets which are being harmonized in various ways.

From a political and a so-called 'managerial' point of view, these various projects have been grouped around the idea of a genuine statute for a European company as endorsed, under Belgian presidency, on 9 October.

Besides the priority objectives on financial services, directly or indirectly attached to the statute of European society, under the Belgian presidency, the Council, after a particularly successful conciliation procedure with the Parliament, adopted the directive amending the directive on prevention of the use of the financial system for the purposes of money laundering. This adoption was aimed for since the start of the Belgian presidency, long before the tragic events of 11 September.

- Last but not least, the Belgian presidency also paid special attention to examining the draft directive on financial guarantee contracts (the so-called 'collateral directive').

The Belgian presidency considered this last draft directive as very important for the future development of the European capital markets. Some lawyers as well as some political decision makers might believe that collateral law is something far too technical and specialized to be interesting, let alone to be politically relevant.

Well, they are seriously mistaken. First, collateral is the absolute condition for the provision of credit. Without collateral, there is simply no lending. Since the economic wealth of a nation very much depends on the health of its credit, one can see that an adequate collateral legislation directly favours economic development, and accordingly creates jobs. Second, collateral is the risk mitigation technique *par excellence*. Especially in the derivatives markets, where counterparties hold massive exposure on each other, collateral is, together with netting, the factor of financial stability without which no economy can thrive.

In retrospect, the results of the Belgian presidency can be summarized as follows. The Council has now succeeded in adopting:

- the directive amending the directive on prevention of the use of the financial system for the purposes of money laundering;
- the two UCITS directives;
- a common position on the proposal for a directive concerning the distance marketing of consumer financial services;
- the proposal for a regulation on cross-border payments in euro;
- a political agreement on the proposal for a directive on insurance intermediaries.

At the end of the Belgian presidency, the Council has also adopted a general approach with regard to three dossiers of special relevance in measuring how successfully the Action Plan has been implemented, namely the proposals for a directive on market abuse, a directive on financial guarantees ('collateral') and a regulation on the application of International Accounting Standards.

A huge amount of preparatory work has also been done on the remaining texts (financial conglomerates, prospectuses and occupational pensions institutions), which should enable future presidencies to proceed on a sound basis. Detailed reports on the progress of work on the proposal for a 'prospectus directive' and a 'financial conglomerates directive' were submitted to the Ecofin Council on 4 December 2001.

One of the significant points in the analysis of the members of the Lamfalussy Committee consists in drawing attention to the fact that

> Market participants and some regulators commented that they would like also to see greatly reinforced transparency, openness, and accountability at all levels. Early and institutionalised involvement of market practitioners and consumers in the legislative process is strongly recommended. Deadlines at every stage of the procedures are widely supported.

Those who speak of transparency, accountability and institutionalized involvement also mention discipline in the negotiations and a proactive attitude on behalf of all concerned parties. The problem of cross-border payments in euro raises some questions. One has to admit that the results of different inquiries by the European Commission are particularly disappointing in regard to the importance of the costs of those transactions. However, from a sociological point of view, it was obvious that the layman would not understand that, from 1 January 2002 onwards, with euro coins in his pockets, cross-border transaction costs in euro would remain at their former level, or even marginally increase, as predicted by a recent study of the European Commission. In this matter, the Internal Market Council has succeeded in adopting a political agreement on (and adoption of) the proposal for a regulation on cross-border payments in euro on the basis of the Commission proposal to align the rates of these payments on national rates.

If the adoption of this regulation by some is considered to be a failure on behalf of different representatives of the banking sector, it should, however, be noted that the consequences for credit institutions must not be exaggerated. Indeed, considering the fact that the number of cross-boarder payments is, at present, very low, we can assume that the impact of the regulation in terms of costs should remain limited. Furthermore, the foreseeable growth of these payments cannot be but very progressive. This will allow a gradual implementation in order to align the necessary structures with the expected costs.

With regard to this issue, at the informal Ecofin Council at Liège, 21 and 22 September, a consensus was reached to request the Central Bank of Europe to present a profound study on the integration of payment systems in the European Union.

In a more proactive way and in terms of positive lessons that might be taken from this experience with regard to 'lobbying', we are convinced that in these matters, the best way to proceed is to create appropriate forms of partnerships among private actors and political representatives, while respecting clear deadlines and firm engagements in order to prevent any 'expectation gap' in the public. This was the approach taken by the Belgian Government when reinforcing the information given to consumers with regard to the passage to the euro. Indeed, long-standing initiatives were reinforced in collaboration with all concerned sectors (conversion of banking accounts in euro, price displays in euro in many gas stations or double displays in warehouses and a growing number of shops).

Inevitably, this European and international context has consequences, in Belgium, on the institutional architecture of the financial markets and on the institutional supervision. The conclusions drawn should address the problems resulting from an increasingly strong competition in respect of adequate deregulations, the qualitative standards of regulated environments and the monitoring of survey systems. This reflection should guarantee the market's efficiency and integrity, in preserving not only the balancing of capital supply and demand, but also in satisfying the needs of issuers, protecting the interests of investors and assuring highly qualitative services by intermediary parties.

In the Belgian context, such a reflection is particularly welcome. As a matter of fact, in 2000, the Parliament has adopted a new legislation permitting not only the privatization of the Brussels stock exchange, combined with its integration in the Euronext structure, but also the important modernization of its way of functioning. Also, the activities of EASDAQ have recently been integrated by NASDAQ Europe.

Complementing a first wave of organizational reforms, a second part of reform aimed at the regulation of the supervision of the Belgian financial markets.

This second reform at the same time implements the principles underlying a regulation equalling the best international standards, harmonizes the supervision structures of the different secondary financial markets (either Euronext Brussels, NASDAQ Europe or the market in Belgian treasury bonds), while respecting the obligations of the applicable European directives.

Considering the recommendations of the Committee of Wise Men, the text of the Belgian reform innovates in organizing an adequate legal ground for:

- the licensing for clearing and settlement bodies;
- bringing into play a set of regulations on the organization and the functioning of the non-regulated markets in financial instruments, the Alternative Trading Systems (ATS) included;
- the adoption of different ways to implement a code of conduct applicable to financial intermediaries, having regard to investment services made for professional investors (or 'sophisticated investors') on the one hand and for other investors on the other hand;
- the possibility for the Belgian supervisory body (*Commission Bancaire et Financière* (CBF)) to conclude different types of cooperation agreements with existing Belgian and foreign organizations and authorities, in order to partly split up its supervision tasks. This could possibly implicate the designation of

one of these authorities as supervising coordinator (competence created by the draft directive on the prudential supervision of financial conglomerates);

- the choice of adequate regulations on the organization, the functioning, the supervision and the control of Belgian markets specialized in Belgian treasury bonds (at the same time respecting a certain coherence vis-à-vis other secondary markets in financial instruments).

Regarding all the above, we can undoubtedly conclude that the contributions of the Belgian presidency to the drafting of directives and regulations and the relations it organized with the other European institutions involved in the new 'Lamfalussy approach', bear witness to its proactive constructiveness with regard to the European integration of financial services and its rendering possible to respond for the first time in operational terms to different challenges posed when regulating these sectors.

Acknowledgement

The authors thank Chris Redant for his contribution to the editing of this paper.

Part II

Contributions on technology and financial institutions

4 Technology and the new economics of banking

A UK perspective*

David T. Llewellyn

Introduction

The focus of the colloquium is how, in various ways, technology is impacting on the financial services industry. The purpose of this overview paper is to offer a perspective from the UK of how technology is changing the underlying economics of the banking industry. Our starting point is simple: technology is changing the fundamental economics of banking, just as it has with many other industries.

The pressures impinging on the financial sector, many of which are technology driven, have the potential to transform the structure of financial systems, the type and range of financial institutions conducting financial services, the type of business undertaken by financial firms, and how business is conducted. All aspects of banking (and financial services) business are being transformed by the impact of technology, and this includes issues related to optimal organizational structures of financial firms.

The context is that the banking and retail financial services industries are in an era of substantial change where a combination of pressures operating simultaneously is changing all aspects of the business. These pressures are having a substantial impact on the structure and intensity of competition. New types of competitors are emerging, some retail financial services markets have become more *contestable*, which institutions conduct what business is changing, the business profiles of individual financial firms are undergoing change as is the way business is conducted (manufacture, origination, processing). Substantial changes are also emerging in the nature of the interface between product providers and consumers. Consolidation in many parts of the financial services industry has accelerated, and the overall structure of the industry is in a period of transformation.

As a result of these pressures, banks around the world face formidable challenges in each of three key areas: strategy, business operations, and organizational structure. In particular, as entry barriers into banking services are eroded, banks are increasingly facing competition from a wider range of actual and potential suppliers of banking services: the capital markets, money markets, non-banking financial institutions, and also ‘non-financial banking institutions’.

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

Within this general context, four central themes are outlined in this paper:

1 A *combination* of pressures operating simultaneously is changing all aspects of banking business in a fundamental way and to an extent which represents a *paradigm shift*.
2 One of these pressures in particular (technology) is a dominant driver of change and a major contributor to changing the economics of the industry.
3 The pressures inducing structural change in the industry are not *incremental* (a continuation of past trends in a steady evolutionary path) but represent a *paradigm shift* where the total impact of the pressures is greater than the sum of the components. A paradigm shift is when the underlying economics of an industry and firms within it change significantly.
4 While many of the pressures operating on the industry can be viewed as potential threats, they also widen (rather than close down) strategic options for firms operating in the industry.

In particular, it is unlikely that a single model will come to dominate the banking industry. It is more likely that the financial system will come to be made up of a plurality of financial firms with major differences in three dimensions: size, product diversification and organizational structure. With respect to the last-mentioned, and in the paradigm of *contract banking* outlined below, this amounts to differences in the mix of internal and external contracts for the provision of processing services.

Pressures producing change

Given the theme of the colloquium, particular emphasis is given to the role of technology in changing the economics of banking and retail financial services. However, this is set in a wider context. There are three generic reasons why current structural and operational change is substantial in the banking and retail financial services industries:

1 There is a powerful set of *combined* pressures operating simultaneously on the industry and which, taken together, have the effect of changing the underlying economics of the industry and the financial firm.
2 One of the dominant pressures (technology) is impacting on the core elements of any financial firm's business: i.e. information, risk assessment, monitoring, processing, consumer interface and delivery.
3 In many areas (both in terms of products and component processes) competition is developing from outside the traditional industry. The evolution of competitive pressures from outside a traditional industry is particularly significant because new entrants are different from incumbents: new entrants have different cost structures, their underlying economics are different, their business models and strategies are different, and they have no interest in following the traditional 'rules of the competitive game' as developed over time by incumbents.

This all amounts to a change in the underlying economics of the banking and retail financial services industries and the economics of firms supplying services within it: what might be termed the *New Economics of Retail Financial Services* (NERFS). The underlying economics of any industry relate to issues such as: the structure and intensity of the competitive environment; the strength of entry barriers and the degree of *contestability* in the industry; the existence or otherwise of economies of scale; optimum organizational structures of firms; the structure of the industry; the way firms and consumers interface with each other; the way business is conducted (origination, manufacture, processing and delivery); and the nature of consumer demand for the products and services of the industry. Our theme is that the pressures on the banking and retail financial services industries are impacting significantly on all of these dimensions simultaneously, and that a dominant driver of change in each of these areas is technology.

The role of technology

As in many industries that have been subject to substantial structural change, technology (in the case of finance, information, trading, processing and delivery technology) is one of the major pressures changing the underlying economics of the industry.

In a recent report, the European Central Bank (1999) offers a comprehensive and succinct analysis of how technology is influencing the banking industry. Its opening sentence sets the scene: 'The developments in information collection, storage, processing, transmission and distribution technologies have influenced and continue to influence all aspects of banking activity.' It goes on to argue: 'Information technology developments have an impact on practically all aspects of banking and can be regarded as one of the main driving forces for change in the banking sector…'. The current paper extends the analysis of the ECB report by arguing that the impact of technology in banking is yet more fundamental and pervasive than suggested by the ECB in that it is changing the fundamental economics of the banking industry and the economics of the banking firm.

In this regard, a distinction is made between what might be termed *managerial technology*, which is incremental in nature and which, while raising internal efficiency in processes, does not change the underlying economics of the industry and can be accommodated within exisiting business models, and *paradigm technology*, which does change underlying structures, business processes and the basic economics of the industry.

There are several reasons why the impact of technology in banking has become especially powerful compared with many other industries, and why this represents something of a paradigm shift:

- It is impacting on the fundamentals of banking business: information, risk analysis, distribution, monitoring and processing. In particular, it enhances management's access to information. Given that banks are ultimately in the 'information business' (Llewellyn, 1999), anything that impacts on the cost, availability and management of information must have a decisive influence on their business.

- It is simultaneously impacting on the production, distribution, delivery, processing and consumer interface processes in banking.
- As argued below, it has the effect of lowering entry barriers and increasing the degree of contestability of banking markets.
- Technology-based decision-making models are emerging for consumers.
- Economies of scale have been raised in bank processes.
- Some aspects of technology have a significant impact on the structure of fixed vs. variable costs. Technology has tended to raise fixed costs but lower variable costs in banking most especially in processing.
- It has impacted on optimal organizational structures of financial firms.
- The interface between consumers and their banks is also changing under the influence of new delivery technology (see also ECB, 1999).

These are discussed in more detail in the remainder of this paper.

The power of technology will be, and has been, decisive: it acts as both a threat and an opportunity to banks. It enables existing services to be provided more efficiently and also enables new services to be offered.

In many ways, banks have lost some of their traditional monopolies and comparative advantages. In particular, a core competence of a bank is the information base which results from managing customers' bank accounts. However, banks are losing some of their traditional information monopolies as information technology has the effect of increasing the supply, and reducing the cost, of information to a wider range of suppliers of financial services. For instance, information technology enables supermarkets to use loyalty cards to build up a profile of customers by analysing their expenditure patterns in the store. This general trend is reinforced by more public disclosure of information by companies and the development of rating agencies which are able to analyse information and assess firms' credit standing.

Technology is transforming the fundamental economics of the retail financial services industry. While the application of technology in the 1970s, 1980s and early 1990s was predominantly *managerial* in nature, current technology is different and represents a *paradigm shift* rather than being incremental in nature.

Some implications of technology

A central theme has been that, combined with other factors operating simultaneously, technology is having a major impact on all aspects of banking business. The remainder of the paper considers six aspects in particular:

1. how technology is influencing competition and the degree of contestability in banking;

2. the impact of technology on economies of scale;
3. the economics of delivery;
4. the emergence of multimedia and how technology alters the nature of the interface between bank customers and their banks;
5. how technology, combined with a process of deconstruction, has implications for optimal organizational structures of banking firms;
6. what the implications of these trends might be for the structure of the financial system and, in particular, whether a single dominant model for the banking firm will emerge.

1 Declining entry barriers and increased contestability

The intensity of competition, and the structure of competitive pressures in the industry, have been changing markedly. Some banking markets (rather than necessarily the banking industry) have become more contestable in that entry and exit barriers have been reducing in significance. The distinction is made because banking products and services can be unbundled and new firms are able to enter some banking markets without offering the full range of traditional banking products and services. Several factors (many of which are technology based) have raised the contestability of banking markets:

- The development of information technology increases the supply, and lowers the cost, of information, and enables new entrants to access and process information.
- The development of technology-based credit-scoring techniques, coupled with greater access to information, enables new entrants to assess credit risks without having the experience gained though managing borrowers' bank accounts over a period of years. This lowers the economies of scope advantages traditionally possessed by banks.
- *Deconstruction* (the unbundling of products and services into their component parts) enables new entrants to compete by subcontracting some of the processes involved in financial services and to effectively buy-in economies of scale from outside. This lowers entry barriers in three ways:
  1. new firms can compete without themselves undertaking all of the processes involved in a particular service;
  2. it enables entry without the requirement of substantial fixed costs which are involved with some processes;
  3. new firms are able to enter without having all of the necessary expertise, as gaps in expertise can be bought in from outside.
- Scale has become less of an entry barrier to the extent that, while technology has increased the economies of scale in processing, many processes can be subcontracted as, with lower fixed costs through subcontracting, economies of scale can effectively be bought-in from specialist providers of processing services. Scale economies tend to be in bank processes rather than in banks *per se* which means that, if processes can be subcontracted, economies of scale can be secured by firms of varying size.

- As new forms (most especially telephonic and computer based) of delivering banking services have emerged and developed rapidly, the branch network (traditionally an entry barrier) has become less significant. New entrants are able to offer banking services without the necessity of an extensive and costly branch network with concomitant heavy fixed costs. This is lowering both entry and exit barriers.
- In some banking markets (notably wholesale lending) the steady globalization of banking markets (itself induced in part by technology) has made local (national) markets increasingly contestable as large borrowers have access to global banking markets.
- The development of Internet facilities for banking products and services has also enhanced the contestability of banking markets. Above all, the Internet means that search costs for consumers and advertising costs for suppliers have been lowered substantially. It also means that distance between supplier and consumer has become less significant.

Recent new entrants into banking markets have certain characteristics in common: they tend to be highly focused within a narrow part of the value chain; they have low fixed costs; they have low or zero legacy costs; they have a (sometimes substantial) franchise value in their brand name; they often enter in partnership with incumbent banks; they tend to use technology-based delivery systems; they have low exit barriers; and are focused in a narrow product range.

2 Economies of scale

A key strategic issue for financial firms relates to whether there are economies of scale and, if so, the origin of such economies. The traditional empirical literature does not yield unambiguous results and tends to conclude that there are only limited economies of scale in banking.

An enormous academic empirical literature has attempted to identify and measure economies and diseconomies of scale in banks. The results are, at best, inconclusive and ambiguous. Overall, except with comparatively small banks, there seems to be little support for the proposition that large banks have lower average costs than smaller banks. Economies of scale seem to be exhausted at comparatively low levels. In fact, there is some evidence that increasing size beyond a certain point has the effect of raising unit costs. The overwhelming conclusion is that the major determinant of a bank's cost level is not size *per se* but its own internal efficiency (X-inefficiency). In other words, the variation in costs between banks of similar size is greater than that between banks of different size.

Many research studies of financial institution scale efficiencies in the late 1980s and early 1990s used US data from the 1980s. The consensus finding was that the average cost curve had a relatively flat U-shape with medium-sized banks being slightly more scale efficient than both large and small banks. Only small banks had the potential for economically efficient scale efficiency gains. The location of the lowest point on the average cost curve differed between studies but was usually between about \$100 million and \$10 billion in assets. For a review of the empirical evidence see: Hunter and Timme (1986); Berger *et al.* (1987); Berger *et al.* (1999); Ferrier and Lovell (1990); Hunter *et al.* (1990); Noulas *et al.* (1990); Berger and Humphrey (1991); Mester (1992); Bauer *et al.* (1993) and Clark (1996). Although there were differences between the studies, almost all suggested there were no significant scale efficiencies to be gained, and possibly some slight scale efficiency losses to be suffered, from mergers and acquisitions involving large banks. As put by Bisignano (1998) in the context of a wave of bank mergers: 'with reference to banks a puzzle exists, since there is little empirical analysis which confirms the existence of significant economies of scale in banking … the merger-acquisition wave in banking is difficult to understand'. On the other hand, there are motives other than scale that also drive bank mergers and acquisitions.

There is substantial evidence that X-inefficiency dominates over scale inefficiency. Overall, there is no firm evidence that size *per se* is a necessary ingredient of success, or a source of sustained competitive advantage.

However, more recent literature (applying different statistical methodologies) is tending in the direction of identifying greater economies of scale than in the past. This research, which applies different econometric techniques, different efficiency concepts, and/or more recent data from the 1990s, suggests that there may be more substantial scale, scope and product-mix efficiency gains available from consolidation. For instance, Berger and Mester (1997) found that 1990s data displayed substantial cost-scale economies of the order of about 20 per cent of costs, for bank sizes up to about \$10 to \$25 billion of assets. It would appear that the potential for scale economies has increased in the 1990s compared with earlier periods.

In some areas of banking, modern technology is itself raising the economies of scale. However, this too is ambiguous as a distinction needs to be made between economies of scale in banks as opposed to bank processes. One of the factors raising economies of scale in banking is the impact of new information, processing, trading and delivery technology. New technology is often very expensive to install, and the advantage of being big is that the full economic use of new technology is often only feasible on the basis of a very large scale of operations. The cost of technology is very high and has become a major aspect of the cost structure of banks. In this regard, some banks cite the cost of installing new technology as a major motive for a merger.

When considering economies of scale, a distinction needs to be made between banks and bank processes. Individual bank processes (cheque clearing, credit card administration, etc.) do show economies of scale. Combining the two conclusions seems to suggest that there are no clear economies of scale in banks (in that large banks do not consistently have lower average costs than smaller banks) but they do exist in banking.

This is an important conclusion as, while competitive pressures are forcing banks to lower their costs, there are other ways of achieving economies of scale in bank processing other than by being a big bank. A central strategic issue for financial firms of all sizes is how, therefore, to secure the competitive imperatives of economies of scale. However, in the NERFS being big is only one way of securing such economies of scale: other options include establishing joint ventures, outsourcing processes, insourcing processes, or creating a confederation of financial firms. In effect, small firms which are unable to secure economies of scale internally are able to secure them externally, i.e. by buying into economies of scale externally.

3 Delivery systems

A major strategic issue being addressed by all financial services firms is the role of technology in changing the economics of delivering financial services (Howcroft, 1987). Technology has a major impact on the way banking and financial services are delivered. In particular, it reduces the dependence on the branch network as a core delivery mechanism. In this respect, what historically has been one of the banks' major competitive advantages (the branch network which acted as an effective entry barrier) may have become one of their most difficult problems. This is because a significant part of the cost structure of a bank is determined by the basic infrastructure. With the development of new technology, a wide range of alternative delivery mechanisms becomes available and most especially through electronic media: ATMs, fixed and wireless telephone, home banking, interactive television, proprietary PC-based services, interactive multimedia kiosks, the Internet, etc. In addition, new types of branches are emerging: machine-based, video conferencing with access to staff, 'banking malls', multimedia kiosks, etc. Their location is also changing as some banks seek to place some types of branch in strategic locations such as retail stores with a substantial customer throughput.

Financial services firms are having to offer a wide range of access routes. However, adding new distribution channels adds another layer of costs while not in itself removing the overheads of the branch network.

One particular aspect of the application of technology in banking, and how the economics of the industry are changing, relates to the Internet. The potential impact of the Internet on the banking and retail financial services industries is substantial:

- the marginal cost of transactions is virtually zero;
- distance between consumer and supplier becomes meaningless and of no economic significance – this may result in more cross-border competition;
- the consumer pays the access costs;
- as an increasing number of rival banks and financial firms open net sites and home pages, the cost of information to the consumer and the search costs for rival services and products become very low which in itself increases competitive pressures in the market;
- the transactions costs of switching between competitors are reduced which is likely to have the effect of eroding customer loyalty;
- it further erodes the necessity to have a branch network to supply financial services and further erodes entry barriers.

New and emerging technology is having a significant impact on the relationship between service suppliers (financial services firms) and their customers. In essence it changes the structure of the intermediation between the supplier and the customer. At one level, financial institutions are becoming less remote from their customers who, through new forms of delivery, gain easier and more immediate access to the firm and at times and locations which are at the customers' choice. In the opposite direction, the firm also has easier and quicker access to online information about the customer and the available range of products on offer by itself and its competitors.

However, of some concern to banks, the relationship between service supplier and the customer is also becoming more remote through new technology. First, technology often replaces service through immediate access to staff and customers may feel that this creates a more remote relationship. Second, the impact of new technology tends to raise the number of firms in the value chain between the customer and the ultimate product supplier. Third, some of the companies within this value chain may themselves become competitors to the supplier.

What seems to be emerging is that banks are developing delivery matrices with differentiations made both between different products and services on the one hand, and different customer groups on the other. Banks now offer a choice of access routes. Thus, a given service will be offered to different customer groups through a range of alternative delivery channels, and individual customers will use a range of alternative delivery mechanisms for different services and products. Choice in delivery is emerging as a key element in successful competitive strategy. However, this is likely to be expensive as, to allow for customer choice, excess capacity may be needed in each delivery mode. This in turn is likely to lead to the explicit charging for different delivery mechanisms.

Developments in technology mean that financial systems are substantially over-supplied with infrastructure and overlapping delivery systems through a duplication of branch networks. Delivery strategies are evolving at two levels: branch networks are being rationalized, and banks are widening the range of delivery options.

4 Multimedia and consumer interface

A particular dimension to the impact of technology in financial services focuses on multimedia. In a recent study of the potential of multimedia in banking, the PA Consulting Group argues that there are five key roles to be performed in the multimedia value chain and that banks, as the ultimate product providers, stand at the end of a queue of relationships with the customer. These five roles may be summarized as follows:

- The customers' immediate point of contact is with a device supplier which provides the customer with the boxes that enable him or her to access and navigate online services: they provide appropriate hardware and software to the customer to facilitate access to services.

- Network operators (such as telephone, cable or television companies) control the infrastructure needed to deliver services to customers.
- Service managers (such as Microsoft) provide the operating systems through which products are sold and billed; they provide the management and technical environment to enable services to be marketed to customers, and they are able to provide customer management services.
- Service providers can package specific materials into service bundles capable of being marketed to customers.
- Content providers create the specific material for consumption by customers.

Such a picture is a radical, though readily feasible, departure from the traditional view of an integrated financial institution providing services and products to its customers with which it has a dedicated link without the intermediation of other companies in the value chain.

A large part of the value added in financial services is now provided not by the product provider itself but by intermediating firms in the process. In this new environment, device suppliers, network operators and service managers have direct contact with the customer, and consumers' access to the product provider is through these intermediaries. An increasing number of financial services providers, including new entrants such as supermarkets in the UK, are joining forces with non-financial organizations (sometimes computer companies) in strategic alliances. The rationale is to work together on the basis of complementary core competences and the economies of scale to be derived through specializing in areas where core competences are most powerful.

There is no immediate threat in this structure to the relationship between customers and product providers, providing the intermediate companies in the value chain are viewed exclusively as facilitators and not competitors to the product provider. However, some of the technology being supplied by the intermediating agents is interactive in nature and, in addition to controlling the networks, some of the companies have the potential to also become product providers themselves.

In effect, what might begin as a cooperative venture between a product provider and an intermediary, in time may become a means through which the product provider gives access to its customer base to new competitors. If cable and telephone companies control the network used to deliver financial services to customers it might be a small step to also become a product provider by supplying its own financial products and services.

5 Organization structure of the banking firm

Technology is also changing optimal organizational structures for financial firms though in different ways for different firms. The traditional structure of a financial firm is of joint-production (vertical integration) with the firm undertaking all aspects of the business from origination to processing. The *deconstruction* process (whereby financial products are deconstructed into their component parts: origination, manufacture, administration, processing, etc.) focuses on what might be termed *contract banking* which implies financial firms creating internal and external markets for processes. Some processes may optimally be undertaken internally while others are subcontracted. In the contract banking model, the firm has two sets of contracts: between itself and customers (service standards, price, etc.) and also with internal and external suppliers of the components that make up a financial product or service.

This can be considered by reference to a standard bank loan (such as a mortgage) which can be decomposed into three main components: origination, management and asset-holding. A loan has to be originated (a borrower located), subsequently administered (processed) and held on a balance sheet. This has traditionally been undertaken as a single process by a lending institution. And yet different agents may have comparative advantages in different parts of the process and there is no necessary presumption that a single institution is the most efficient at undertaking all aspects of the loan. Thus a bank may have an advantage in originating loans (e.g. through the branch network) and yet not have the most efficient processing facilities. In that case a bank can originate loans the processing of which, on the payment of a fee, is undertaken by other institutions which have a comparative advantage (perhaps because of economies of scale) in this particular activity.

The paradigm of contract banking

Thus technology, and the ability to *deconstruct*, have implications for the optimal structure of the banking firm. Traditional banking involves a joint production technology that produces deposit, lending and transactions services within a given institution. This structure has faced an increasing challenge from separate production technologies. In this way technology can change the underlying economics of the financial firm in terms of its optimal organizational structure.

This process of deconstruction and outsourcing, which is common in the manufacture of goods, has not been the norm in banking where traditionally the banking firm has offered an integrated service by itself providing the service and their components. However, as already noted, the process of deconstruction changes this picture. It enables particular subcomponents of products or services to be subcontracted (outsourced) and supplied by other firms on a contract basis. Similarly, deconstruction enables a bank to provide a particular subcomponent of a service to competitors. Thus, a bank may subcontract the administration of its credit card operation while at the same time export to other banks its risk analysis capacity. The potential exists because the economies of scale in bank processes vary. By subcontracting a particular process, a small bank may be able to buy into economies of scale that it could not itself achieve.

Thus the banking firm can be viewed as a firm which has an interface with a customer base (supplying a range of apparently integrated products and services) and demanding a series of support services in order to supply the services. A distinction is made between the final products and services that the customer demands and the bank supplies (e.g. loans), and the various components of each product or service (e.g. risk analysis, administration, etc.).

The central issue is which of the components are to be supplied internally, which are to be subcontracted, and which are to be exported. Core competences of particular banks are relevant in this. Thus, what may appear to a customer as an integrated product or service is in fact a series of deconstructed components which may or may not be supplied from within the bank. The bank defines the components and decides which are to be supplied internally and which subcontracted. In effect, a series of contracts is established by the contracting bank with internal and external suppliers.

Contract banking implies a bank offering a full range of services but where the bank coordinates inputs from a wide range of different companies. The core is a contract the bank has with its customers to supply a set of services or products of a particular standard. In turn, the bank contractor has a set of contracts with a range of internal and external suppliers of the components of these ultimate products and services. The value added by the bank contractor is in the management of these contracts. The concept of contract banking is discussed more fully in Llewellyn (1997).

Various forms of outsourcing are available: third-party processors, service bureaux and facilities management contractors. In addition, two or more parties might establish a joint venture to undertake certain activities on a joint basis.

A major form of outsourcing is third-party processing on a line-of-business basis. Those functions that are most automated or specialized tend to be the most outsourced partly because this is where economies of scale potential are greatest. As already noted, developments in information technology have lowered the cost of performing information-intensive activities, providing economies of scale can be reaped. Examples of outsourcing include: mortgage processing, credit card administration, cheque processing, network operations and management, credit card issuance, student loan processing, trust processing, securities safekeeping, ATM driving/switching, retail lockbox, Applications Development and Management, data centre and balance reporting. As technology becomes more intense and specialized and requires heavy investment, it tends to be disaggregated, i.e. technology operations are broken apart and split up amongst a number of highly specialized technology companies which supply similar services to several banks.

There are several reasons why outsourcing is undertaken and why it has become an increasingly common feature in banking:

- to reap economies of scale that cannot be obtained internally;
- some areas may be too specialized to be undertaken internally;
- a particular expertise may not be available internally and may be uneconomic to acquire;
- to gain increased flexibility in the use of technology;
- to spread (most especially fixed) costs and risks;
- to break an internal monopoly when services are supplied exclusively internally;
- to change the cost structure: lower fixed costs.

Above all, a major advantage of outsourcing is that it transforms fixed costs into variable costs and hence reduces the requirement for, often large, up-front costs in developing and adapting processing facilities. If a firm conducts its own processing it pays, and must recoup, through the pricing of its products and services, both the large fixed and small variable costs of the process. On the other hand, if it subcontracts the process it pays the supplier a proportion of the supplier's fixed costs plus the variable costs. The whole procedure is economic if the higher variable and transactions costs through outsourcing are less than the savings on fixed costs. In addition, the outsourcing firm may find it economic to outsource even when this condition is not met because of the sharing of risks and the greater flexibility it secures through minimizing its infrastructure and fixed costs. In general, firms with low fixed costs and capacity are more flexible than those with high fixed costs even if the variable costs are higher.

If there are significant economies of scale in a particular process a bank can secure these economies in one of four ways: by being big; by outsourcing; by forming joint ventures with others; or by the bank investing in a process and supplying the excess capacity to others. As an example of the last-mentioned, a bank may decide to establish a cheque processing facility where the economies of scale are considerably bigger than the bank itself and to provide the service to other banks.

In a competitive market all firms are under pressure to gain cost advantages wherever they can be secured. In some cases banks may have gone as far as they can in cutting costs without a more fundamental re-engineering of the business such as is implied in contract banking. If technology has the effect of increasing the economies of scale, the issue becomes how banks can reap such economies. As noted, economies can be secured either internally or externally but in some cases it may require a fundamental re-engineering of the bank. However, paradoxically, the technology which increases the economies of scale in bank processes, combined with the ability to deconstruct products and services and have components priced and supplied independently, means that both small and large banks can coexist, and that there will be greater variety in the structure of banking firms. This is because economies of scale are in bank processes rather than in banks *per se*.

At its extreme, the *virtual bank* emerges. This has an interface with its customers and seemingly supplies a set of integrated services and products. And yet it does nothing itself other than manage a set of contracts with external suppliers. It is a contract buyer of other firms' products and services and a coordinator of a network of contracts and services. It is, in effect, a broker between the customer and the ultimate supplier of services which go to make up the final products and services demanded by the customer. This may mean that comparatively small virtual banks can exist alongside large banks. They may provide the full range of banking services with the customer being unaware that the bank is in truth a network of alliances with specialist providers.

6 The structure of the financial system

The development of outsourcing means that there can be a role for the small bank in a market and technology environment where many banking operations require large scale to be economic. While there is a trend towards more consolidation in the banking industry, there will still be a place for the smaller bank though it will not have the traditional structure.

An implication of much of the analysis of this paper is that banks will be under constant pressure to cut and contain costs as a permanent feature of strategy. The economies of scale to be derived through the application of technology will be one of the routes of this pressure. However, if economies of scale relate predominantly to bank processes rather than institutions, and external contracts can be managed efficiently, the existence of economies of scale does not mean that only large banks can be competitive and survive.

What in practice is likely to emerge is a spectrum of different types of bank (Figure 4.1). At one end of the spectrum will be the traditional fully integrated bank which, because of the economies of scale in bank processes, will be very large. At the other end of the spectrum will lie the virtual bank. In practice, the majority of banks will lie within the polar boundaries of the spectrum, with some services being provided internally and others outsourced. The differences between banks, and their location along the spectrum, is ultimately a question of the balance between internal and external contracts, and many alternative structures are likely to emerge.

Financial firms will come to be distinguished in three dimensions: size, product range, and organizational structure (the balance between internal and external contracts).

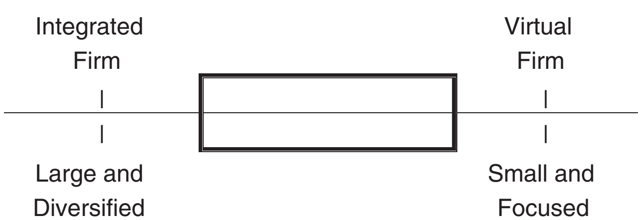


Figure 4.1 Institutional spectrum.

Economies of scale again

These developments bring into question the traditional measures and concepts of economies of scale in banking. If there are alternative ways of securing economies of scale in processes by changing the organizational structure of the bank (e.g. by outsourcing), the same economies of scale can be secured by banks of very different size. Under some extreme assumptions (no external contracting costs, perfect divisibility of processing, etc.) the long-run average cost curve is horizontal and the same for all banks, though different banks will reach this position through different combinations of internal and external contracts for processing, etc. This is illustrated in Figure 4.2 where AC1 represents the average cost curve of a bank with old technology with only internal contracts. AC2 represents the cost curve with new technology with only internal contracts for processes. This indicates that technology is having the effect of increasing the economies of scale, raising the optimum scale, and raising costs at low levels of output compared with old technology. However, if banks have the option of choosing a mix of internal and external contracts (in the latter case they effectively buy-in economies of scale from outside which they cannot secure internally) the average cost curve becomes AC3. In effect, banks of various sizes can secure the same economies of scale. Thus empirical testing might show banks of various sizes having the same costs which might be misinterpreted as there being no economies of scale in banking. The true interpretation is that economies of scale can be secured by small banks through greater use of external contracts. However, this will not be identified with traditional methodologies for measuring economies of scale in banking.

Strategic options

As a result of this, the NERFS is widening, rather than narrowing, the strategic options available to financial firms of all types and sizes. As different financial firms adopt different strategies, the future structure of the retail financial services

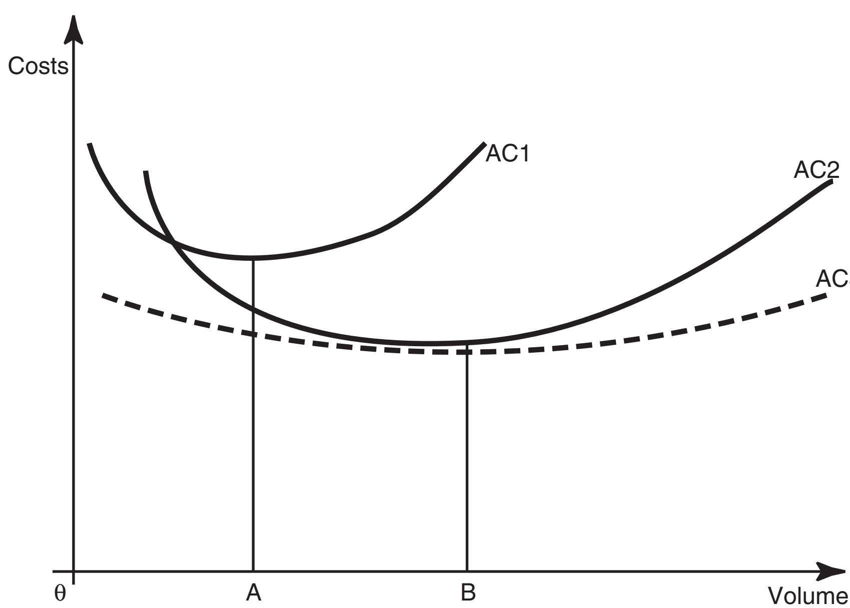


Figure 4.2 Technology and economies of scale.

industry may involve greater diversity of financial firms. There is no reason why the industry is likely to converge on a single, or dominant, model.

With respect to the optimum size of financial firms, there are conflicting pressures: technology is increasing the economies of scale and hence, *prima facie*, a need to be large to gain competitive advantage in the market, but the same technology is making *deconstruction* of processes more feasible and also widening the range of options for securing economies of scale.

There is a viable role for comparatively small and focused financial firms. However, and partly in order to secure economies of scale, success for small financial firms may require major changes in organizational structure at two levels:

1 a radical *deconstruction* of processes with firms focusing upon their core competences and subcontracting many processes which have traditionally been undertaken 'in-house';
2 the formation of confederations of small financial firms to undertake processing collectively for its members while each member retains its competitive independence.

In many European countries and also in Australia, collective arrangements are common most especially in the mutual and cooperative sectors of the financial system. Examples are Federcasse in Italy and the Credit Unions Services Corporation in Australia. All are designed to enable small financial institutions to gain the competitive imperative of economies of scale.

Conclusion

As the underlying economics of the banking and retail financial services industries and firms change in the various dimensions identified, past paradigms of analysis may become less appropriate. In particular, no single dominant business model is likely to emerge and the financial system is likely to come to be populated by a greater diversity of financial firms with respect to size, product mix and organizational structure. This is one of the implications of new technology in banking.

The evolution of national banking systems, and the business of banks in particular countries, are always and everywhere influenced by a combination of *country-specific* and *global* pressures. In the years ahead the relative role of these two sets of forces is likely to change with global pressures becoming more decisive than country-specific factors. This is partly because the dominant pressures (e.g. technology) are themselves global in nature.

In practice, the timing, speed and intensity of the pressures vary from country to country and in some countries regulation continues to offer a degree of protection to the value of the banking franchise. Nevertheless, as competition becomes increasingly global in nature, and many of the pressures (e.g. technology) are universal, no nationality of banks will be immune from the pressures operating on the banking industry.

Opinions differ as to whether the evolution of information, processing, trading and delivery technology represents a process of further incremental change in banking or whether it is truly a paradigm shift. The theme here has been that it is closer to the latter than the former both because the pace of change has accelerated and because it is transforming all aspects of banking business simultaneously.

References

Bauer, P.W. *et al.* (1993) 'Efficiency and productivity growth in US banking', in H.O. Fried *et al.* (eds), *The Measurement of Productive Efficiency: Techniques and Applications.* Oxford: Oxford University Press, 386–413.

Berger, A.N., Hancock, G.A. and Humphrey, D.B. (1987) 'Competitive viability in banking: scale, scope and product mix economies', *Journal of Monetary Economics* 20: 501–20.

Berger, A.N., Demsetz, R.S. and Stratham, P. (1999) 'The consolidation of the financial services industry: causes, consequences and implications for the future', *Journal of Banking and Finance*, (December).

Berger, A.N. and Humphrey, D.B. (1991) 'The dominance of inefficiencies over scale and product mix economies in banking', *Journal of Monetary Economics* 28: 117–48.

Berger, A.N. and Mester, L.J. (1997) 'Inside the black box: what explains differences in the efficiencies of financial institutions?', *Journal of Banking and Finance* 21: 895–947.

Bisignano, J. (1998) 'Towards an understanding of the changing structure of financial intermediation', *SUERF Studies* no. 4, Amsterdam: SUERF.

Clark, J. (1996) 'Economic cost, scale efficiency and competitive viability in banking', *Journal of Money, Credit and Banking* 28: 342–64.

European Central Bank (1999) *Technology and European Banking.* Brussels: ECB.

Ferrier, G. and Lovell, C.A.K. (1990) 'Measuring cost efficiency in banking: econometric and linear programming evidence', *Journal of Econometrics* 46: 229–45.

Howcroft, B. (1987) 'Retail banking in the UK: a change in organisational form', *Journal of Retail Banking* 9(3), 35–42.

Hunter, W.C., Timms, S.G. and Yang, W. (1990) 'An examination of cost subaditivity and multiproduct production in large US banks', *Journal of Money, Credit and Banking* 18: 152–66.

Hunter W.C. and Timme, S.G. (1986) 'Technical change, organizational form, and the structure of bank productivity', *Journal of Money, Credit and Banking* 18: 152–66.

Llewellyn, D.T. (1997) 'Contract banking and its implications for building societies', project paper no. 5, *Economics of Mutuality and the Future of Building Societies.* London: Building Societies Association.

Llewellyn, D.T. (1999) 'The new economics of banking', *SUERF Studies* no. 5, Amsterdam: SUERF.

Mester, L.J. (1992) 'Traditional and nontraditional banking: an information–theoretic approach', *Journal of Banking and Finance* 16: 545–66.

Noulas, A., Ray, S.R. and Miller, S.M. (1990) 'Returns to scale and input substitution for large US banks', *Journal of Money, Credit and Banking* 22: 94–108.

5 The effects of technology on the costs and risks of Spanish banks*

Ignacio Fuentes Egusquiza and Teresa Sastre de Miguel

Introduction

Recent advances in information and communication technology have modified the ways in which customers have access to banks' services and products, leading to a new model of banking (PC and Internet banking, telephone banking) which represents a major change in relation to the most traditional model, mainly based in branch networks. Yet this is no more than the last wave of technological advances which are being incorporated in bank activity since several decades ago. Information processing and storage by electronic means, electronic trading and settlement systems, electronic means of payment and automated teller machines are a few examples of this process. The new technologies make possible cost saving transformations in the production process which may improve efficiency in data management and risk control. Together with these potential benefits there are some other more uncertain effects like some changes in the nature and balance of risks faced by banking institutions.

The purpose of this paper is to address both issues: the potential impact of technology on bank efficiency, which is examined in Section I, and the increasing complexity of risk control in a framework of recent widespread and fast technological advances, which is discussed in Section II. The main conclusions of the paper are summarized in Section III.

I Bank efficiency: the impact of technology on costs

The technological advances that have occurred in the area of information technology and telecommunications in recent years have had a significant influence on the operation of the world financial sector. This section examines the degree to which Spanish banks have incorporated and adapted to these technological developments, as well as their impact on the structure of costs.

In the case of Spain, the process of incorporation of computerized media into the management of banking operations has been closely linked to the adaptation of operating procedures required by the launch of the single currency and the year

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

2000 problem. Although computerized media had been coming into general use at the end of the 1980s, the start of Stage Three of Economic and Monetary Union involved an additional boost to the process, especially in departments related to the activity in the interbank market and in securities markets. The adaptation to the new requirements of Stage Three of EMU involved not merely the adjustment of existing equipment and procedures, but also widespread renewal of the same, with incorporation of the most recent technical advances.

During the lead up to Stage Three, spending on information technology and telecommunications by a large number of banks, especially the commercial banks, increased significantly. This can be seen in the various indicators presented in Table 5.1. Spanish commercial banks tended to increase their IT and communications spending from 1996. However, this process seems to have occurred with a different pattern and intensity in the other Spanish deposit-money institutions. In the case of savings banks and credit cooperatives the growth of IT spending began somewhat later and was concentrated in a shorter time period (1998–9).

The differences in the patterns of behaviour of these three kinds of bank seem to be related to the differences in their product specialization and to the existence of a sectoral institution providing services to the institutions on the basis of cooperation agreements (CECA, Banco Cooperativo[1]). In the case of the commercial banks, especially the large ones, which have significant activity in the interbank and government debt markets, the advent of the single currency required an adaptation effort in these areas, which were those most immediately affected. This type of activity has less weight in the savings bank sector, which is more orientated towards retail banking and was therefore able to adapt more gradually. Also, the savings banks, and particularly the credit cooperatives, channelled a large part of their operations in these markets through CECA or the Banco Cooperativo, which perform interbank and securities management for many of the institutions in both sectors, so that the work to adapt to the single currency and the year 2000 problem was smaller in these sectors.

During the past two years there has been a certain shift in spending from direct spending on applications and computer systems towards the outsourcing of these services. Thus, if only this direct spending is included, the growth in the cost of IT and communications products and services is less than when outsourced services are included (see Table 5.1).

The growth of spending on items related to the incorporation of information technologies corresponds, to a degree difficult to specify, to investment in technology and the incorporation of technical progress in the productive process. The investment in R&D that this represents might be expected to have an impact on average costs and on the cost structure of banks, which may be manifest in different ways.

First, the replacement of labour-intensive procedures with cheaper automated procedures will tend to reduce the cost per transaction. At the same time, the reduction in the number of branches and the associated labour that may result

Table 5.1 Indicators of information technology spending (%) (a)

	Big banks			*Commercial banks*		
	1992–5	*1996–8*	*1999–2000*	*1992–5*	*1996–8*	*1999–2000*
IT expenses/ operating expenses	9.6	11.1	13.8	10.6	11.6	14.7
IT expenses/ overheads	30.5	34.0	41.5	32.0	34.0	41.7
IT expenses/ number of employees (EUR thousands)	5.5	7.8	10.8	6.2	8.1	11.6
IT expenses/ number of branches (EUR thousands)	51.7	64.3	91.6	53.2	62.5	90.3
ATMs (annual rate of growth)	16.8	16.7	12.7	15.1	15.7	11.3
IT expenses (annual rate of growth)	–1.4	17.2	5.7	–1.4	13.0	7.6
IT expenses (annual rate of growth) (b)	–1.4	17.2	–2.8	–1.4	13.0	–1.9
	Savings banks			*Credit cooperatives*		
	1992–5	*1996–8*	*1999–2000*	*1992–5*	*1996–8*	*1999–2000*
IT expenses/ operating expenses	14.0	13.0	15.3	14.3	13.0	14.4
IT expenses/ overheads	36.6	33.8	40.0	36.9	33.1	36.8
IT expenses/ number of employees (EUR thousands)	8.9	9.6	12.2	7.0	7.4	8.7
IT expenses/ number of branches (EUR thousands)	51.6	52.2	64.3	25.4	27.3	32.5
ATMs (annual rate of growth)	8.5	11.1	9.4	93.6	11.1	12.3
IT expenses (annual rate of growth)	3.5	5.9	13.1	3.5	9.2	10.3
IT expenses (annual rate of growth) (b)	3.5	5.9	7.3	3.5	9.2	–4.3

Notes: (a) Information and technology costs also include outsourcing costs in the period 1999–2000; (b) Outsourcing costs are not included in this item.

from the development of different kinds of remote banking (telephone and Internet banking and the use of automated teller machines capable of performing a broad range of bank transactions), may also tend to lead to a shift towards more capital-intensive technologies, thereby reducing average costs and the weight of personnel costs in total operating costs. Other kinds of effect may also occur which will eventually have an impact on the long-run average cost curve. Among these may be highlighted:

- the possibility of gaining access to and harnessing the economies of scale that characterize automated procedures for carrying out transactions;
- a possible rise in the cost of the labour factor, owing to the higher-skilled workers required by the investment in technology who have to be remunerated with higher salaries;
- investment in R&D will tend to lower average fixed costs, in so far as it allows processes to be rationalized and leads to an increase in capital productivity.

However, in the short run there may be fluctuations in average costs arising from an adjustment process that may be different for each of the various variables affected. Also, the costs of banks may be influenced by other factors, other than technological advances, which push them up (e.g. strategies for growth or expansion into new geographical areas). In short, although in the long term one can expect a reduction in average costs, in the short term it may be difficult to determine the influence of the new technological developments on costs in banking activity.

Before analysing the costs of Spanish banks it is worth examining some differences that exist between the various institutional groupings in the Spanish banking system, which have had an impact on trends in costs and the cost structure during the 1990s. For this purpose, Table 5.2 shows the average annual growth in the operating costs of these groups, as well as the average growth in output, approximated through two indicators – total assets and total customer credits and deposits – and the average growth of the branch network. It can be seen that the savings banks and credit cooperatives have recorded higher growth in their activity than the commercial banks. At the same time, their branch network and total number of employees have grown, in contrast to the trends in the same variables in the commercial banks. These data reflect the different strategies adopted by these groups in the 1990s: the expansion into geographical areas outside their traditional locality, in the case of savings banks, and the attempt by the commercial banks to rationalize their branch networks in order to reduce the weight of operating costs in their activity and thus respond to the narrowing of margins. These different strategies have led to much higher growth in the operating costs (both personnel costs and overheads) of savings banks and cooperatives than in those of commercial banks. Thus the weight of staff costs on operating expenses has increased in savings banks while it has decreased in commercial banks.

Meanwhile, there was a larger difference between the growth in total assets and that of customer activity in the case of commercial banks than in that of savings banks and credit cooperatives, which shows the greater importance of activity in financial markets in the case of the former.

A final difference between these groups is the contribution of IT and communications spending to the growth of overheads and operating costs. This contribution was high in the case of the commercial banks, as shown by the fact that the average growth of operating costs falls by 0.6 percentage points when this item is excluded.

Table 5.2 Costs and activity: 1992–2000 (annual % rate of growth)

	Commercial banks	*Big banks*	*Rest of banks*	*Savings banks*	*Credit cooperatives*
Number of branches	–1.2	–1.9	–0.4	3.3	3.5
Number of employees	–2.6	–2.7	–2.4	1.8	3.7
Staff expenses	2.7	2.7	2.6	6.7	7.9
Staff expenses (excl. pension payments)	2.3	2.4	2.2	6.1	8.0
Non-staff expenses	4.0	3.7	4.3	6.1	7.9
Non-staff expenses (excl. IT expenses) (a)	2.1	1.7	2.6	5.7	8.3
Operating expenses	2.8	2.7	3.0	6.5	8.1
Operating expenses (excl. IT expenses) (a)	2.2	2.1	2.4	6.4	8.2
Total assets	8.7	9.2	7.8	10.4	12.7
Customer balances (credit + deposits)	6.5	6.6	6.4	10.7	14.6

Source: Bank of Spain.
Notes: (a) Information and technology costs also include outsourcing costs in the period 1999–2000.

By contrast, in the case of savings banks and credit cooperatives, this growth rate hardly changes when IT spending is not included.

In order to obtain an initial approximation of the impact of information technologies on the operating costs of Spanish banks, Figures 5.1 and 5.2 show the trends in operating costs, overheads and personnel costs per branch, as well as their main explanatory factors.

First, over the last five years, the growth of spending on new technology products and services has been higher than that of other overheads in the case of the large and other commercial banks (see Figures 5.1 and 5.2), contributing significantly to the growth of total overheads. Thus the share of overheads in operating expenses increased three percentage points during 1992–2000 in commercial banks. In contrast, the growth of IT spending only exceeded that of other overheads in 1998 and 1999, in savings banks and credit cooperatives (see Figure 5.2), so that the trend in their overheads per branch has been more influenced by items other than those relating to the use of new technologies. The strategy of opening new branches may have influenced this type of effect.

The strong growth in the cost of IT and communications investment and services in small and medium-sized commercial banks in 1999 was notable. This is largely explained by the investment made by one of the largest banks in this group, which has adopted the strategy of offering a wide range of banking products on the Internet. This bank is currently the one to have achieved the largest foothold in Internet banking in Spain.

As regards personnel costs, their growth was dominated by wage growth, although pension payments helped to widen the gap between these two growth

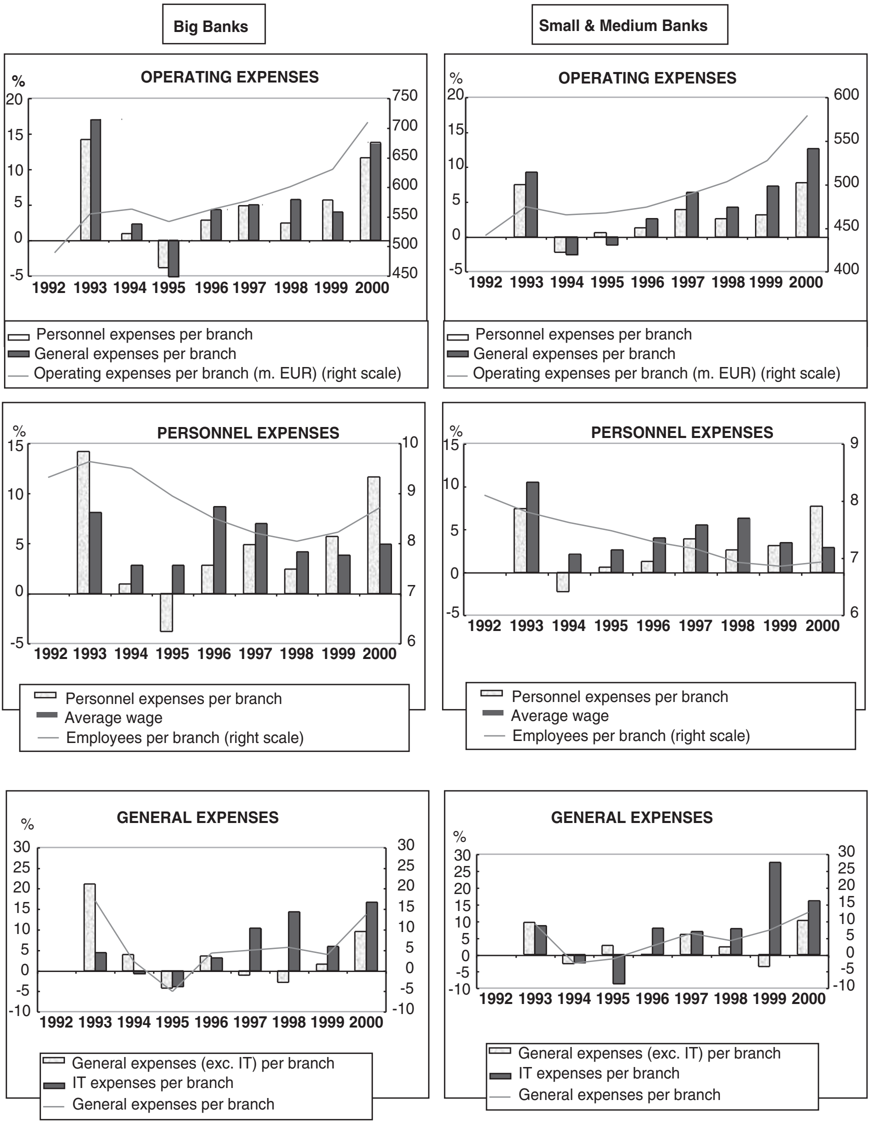


Figure 5.1 Average spending per branch: big, small and medium-sized banks.

rates in those years in which there were a significant number of early retirements or extraordinary contributions were made to pension schemes (as in 2000). The change in the number of employees per branch is another factor explaining the gap between wage growth and growth in personnel costs. In the case of the savings banks and credit cooperatives, the rate of change of labour costs per branch showed a clear downtrend (except in 2000, for the reason

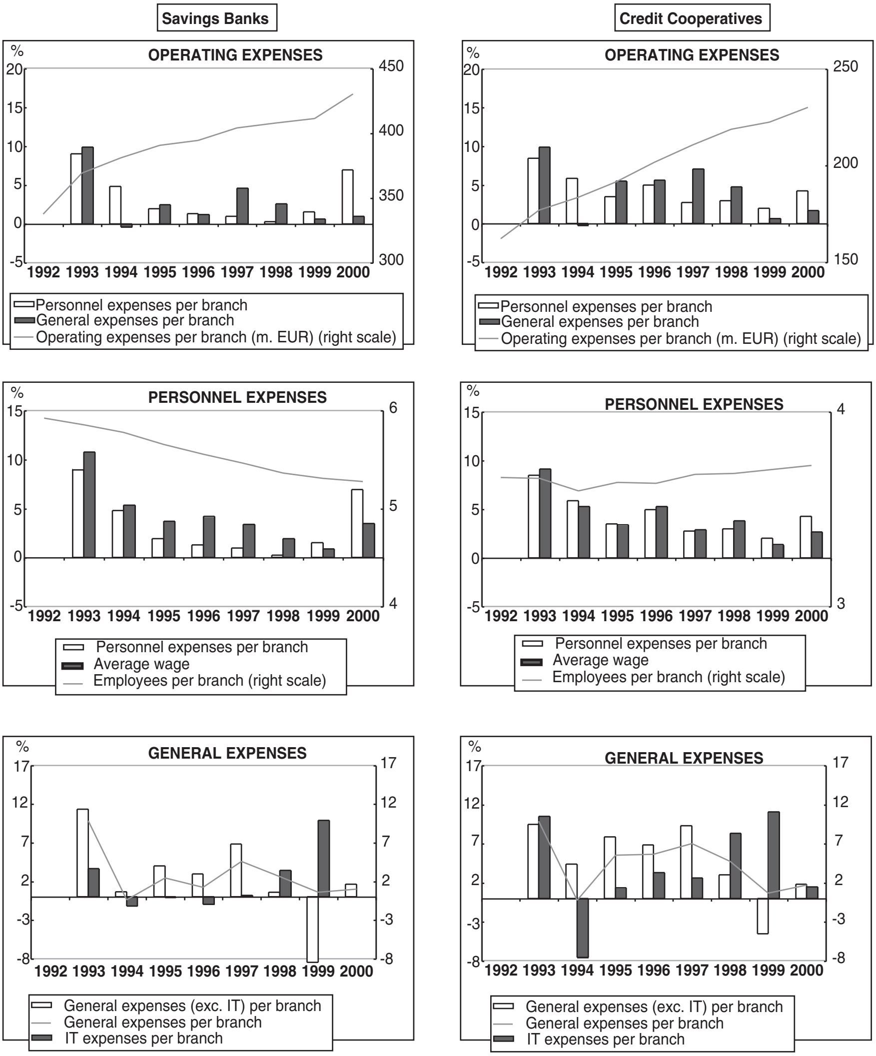


Figure 5.2 Average spending per branch: savings banks and credit cooperatives.

already mentioned), in line with wage growth, which moderated significantly during the second half of the 1990s in the Spanish economy. However, the labour costs of Spanish commercial banks displayed a somewhat different pattern, owing to a larger reduction in the number of employees per branch and to the acceleration of nominal wages during the period 1996–8, the period in which adaptation to the single currency took place and growth in spending on IT systems and processes was higher. This descriptive kind of evidence seems to indicate a possible relationship between technological advances and nominal wage growth, owing to the higher level of skills required by the investment in technology. However, the nature of the data does not allow testing of this hypothesis appropriately due to the small number of observations (which, in some years, are also affected by some known extraordinary events) and their aggregated level which may obscure possible regular patterns if one or two big institutions in the group deviate from those regularities. Yet the increase in the percentage of employees with the status of manager (from 25 per cent in 1995 to 28 per cent in 1999) in savings banks and of those considered technical staff (from 53 per cent in 1995 to 60 per cent in 1999) in commercial banks, would tend to support this hypothesis.

Although the information presented does not enable conclusions to be drawn, it does seem to indicate that, among Spanish commercial banks, there has been a predominant tendency to reduce employment and invest in technology, which tends to improve the technical quality of capital, while, among the savings banks and credit cooperatives, the dominant strategy has been to increase more labour-intensive distribution channels, such as the branch network. The results of these different strategies in terms of efficiency are difficult to assess without resorting to the estimation of cost functions. However, to provide an initial approximation of the effect of these two strategies, some standard ratios used to analyse efficiency in the banking industry are set out in Figure 5.3.

Operating costs per unit of gross income were very stable in the case of both savings banks and credit cooperatives. In the case of commercial banks this indicator was also relatively stable, if the years at the beginning and end of the period considered, when there was unusual growth in gross income, are excluded. According to this indicator, the most efficient group of banks during these years would have been the credit cooperatives. However, the stability of this ratio over time would seem to indicate that it is a target variable for the banks (it is used by rating agencies and supervisory bodies in their assessment of banks) and, in consequence, may bear little relation to the trend in variables indicative of the strategies mentioned.

The efficiency ratios based on balance sheet items, whether total assets or those relating to activity with customers (credits and deposits), display a downward trend in all groups, indicating an improvement in efficiency. Also, a fact already mentioned above is clearly seen: owing to the importance of their operations in the interbank and government debt markets, the large commercial banks are the most efficient when operating costs and total assets are compared, while, in terms of operations with customers, the savings banks and credit cooperatives are the most efficient.

Average operating costs per transaction should, in principle, reflect the cost saving that may have arisen with the incorporation of new information technologies into the productive process. However, this variable displays an upward trend in all the groups, except small and medium-sized commercial banks. Underlying this upward trend is the process of disintermediation from bank deposits to mutual funds, which characterized most of the period considered. This process dominated the trend in this ratio for Spanish deposit-money institutions.[2]

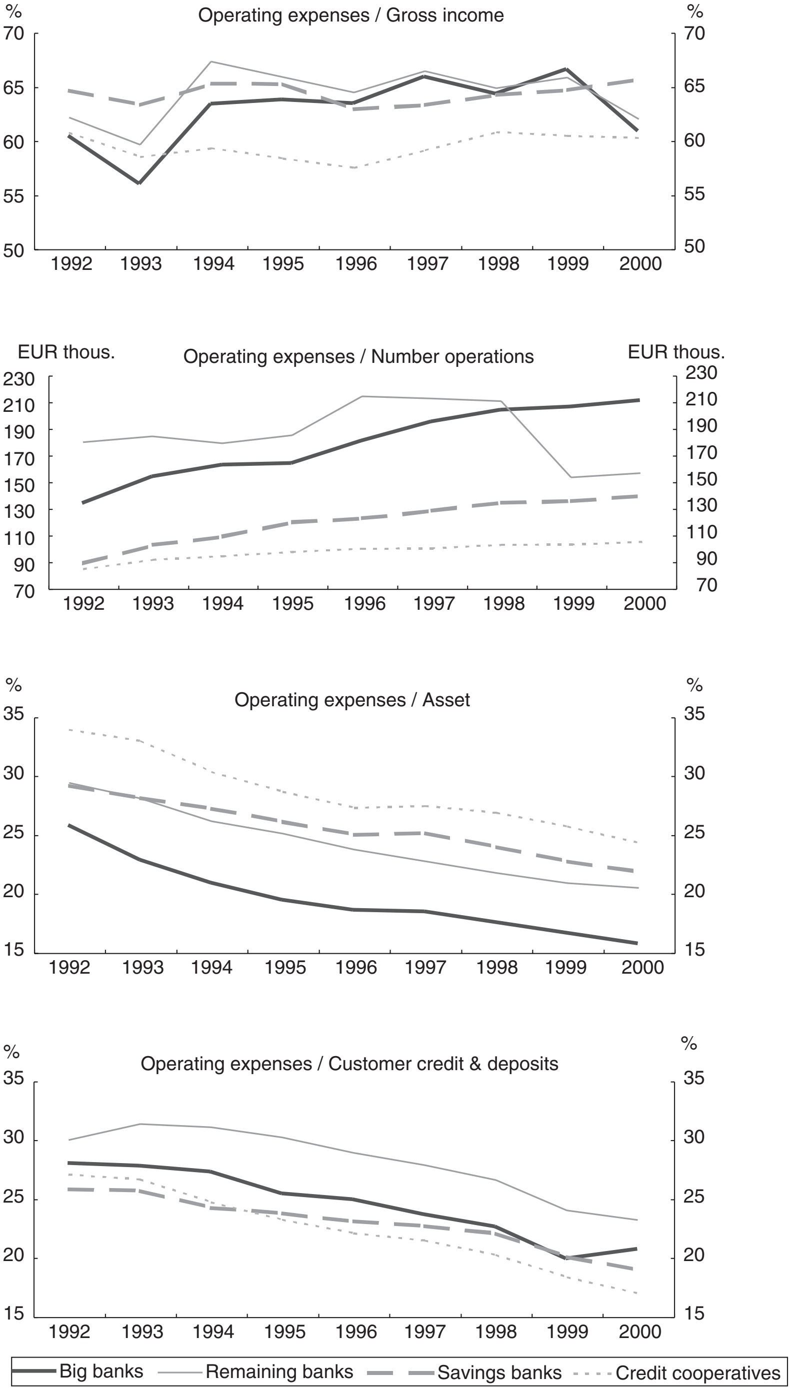


Figure 5.3 Efficiency indicators.

Accordingly, the standard measures of efficiency of an aggregate nature are rather ambiguous and hardly provide any information on the influence of the new technologies on the efficiency of banks, or on the impact of expansion of the most labour-intensive distribution channels implemented by some Spanish banks.

The impact on productivity

Since the cost variables examined provide little information on the impact of information technologies on banks' average costs, this section contains a brief descriptive analysis of labour productivity and its relationship to total factor productivity and spending on technology.

Labour productivity has been approximated by the relationship between total customer credits and deposits and the number of employees.[3] Figure 5.4 shows the rate of change of this variable for four different groups of banks. Since 1994, the growth in labour productivity has tended to rise in Spanish commercial banks, although in 2000 there was a decline in this growth. In the case of savings banks and credit cooperatives, there were also significant improvements in productivity, with somewhat more moderate growth in the middle years of the period considered.

In principle, the improvement in apparent labour productivity in the various groups of banks may be associated with growth in total factor productivity or the capital/labour ratio, or a combination of the two. In Figure 5.4 some evidence of these relationships can be seen in the case of Spanish commercial banks, although their presence seems less clear in the savings bank and credit cooperatives sectors. The capital/labour ratio has been approximated using own funds as a proxy for capital. For its part, total factor productivity has been obtained from the expression:

$$TFP = \frac{Y}{rK + wL}$$

where Y = total customer credits and deposits (in real terms); K = own funds (in real terms); L = number of employees; w and r = market price of the labour and capital factors, respectively.[4]

Among the potential factors that might explain the path of total factor productivity (technical progress, the skill level of workers and, in general, all those aspects that influence the quality of the factors used in the productive process), Figure 5.5 shows the relationship between this variable and the spending by banks on information technologies. In general this type of spending seems to have contributed to the productivity gains seen between 1996 and 1999, although it is among Spanish commercial banks that this relationship seems clearest.

In short, the information of an aggregate nature seems to indicate that the new advances in information technologies may have helped to improve total factor productivity. However, this hypothesis will have to be tested by estimating production or cost functions using data from individual banks in order to avoid biases induced by the aggregation of heterogeneous agents and to increase the number of observations.

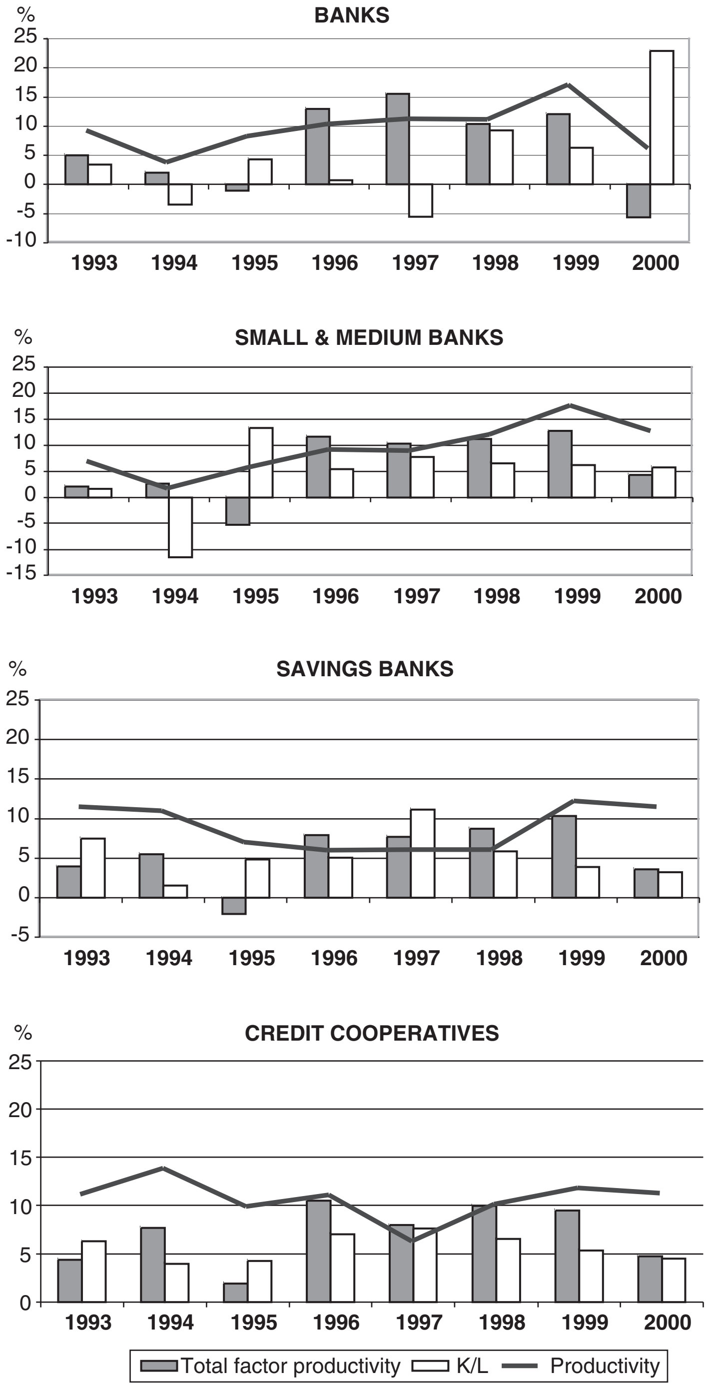


Figure 5.4 Labour productivity (rate of change).

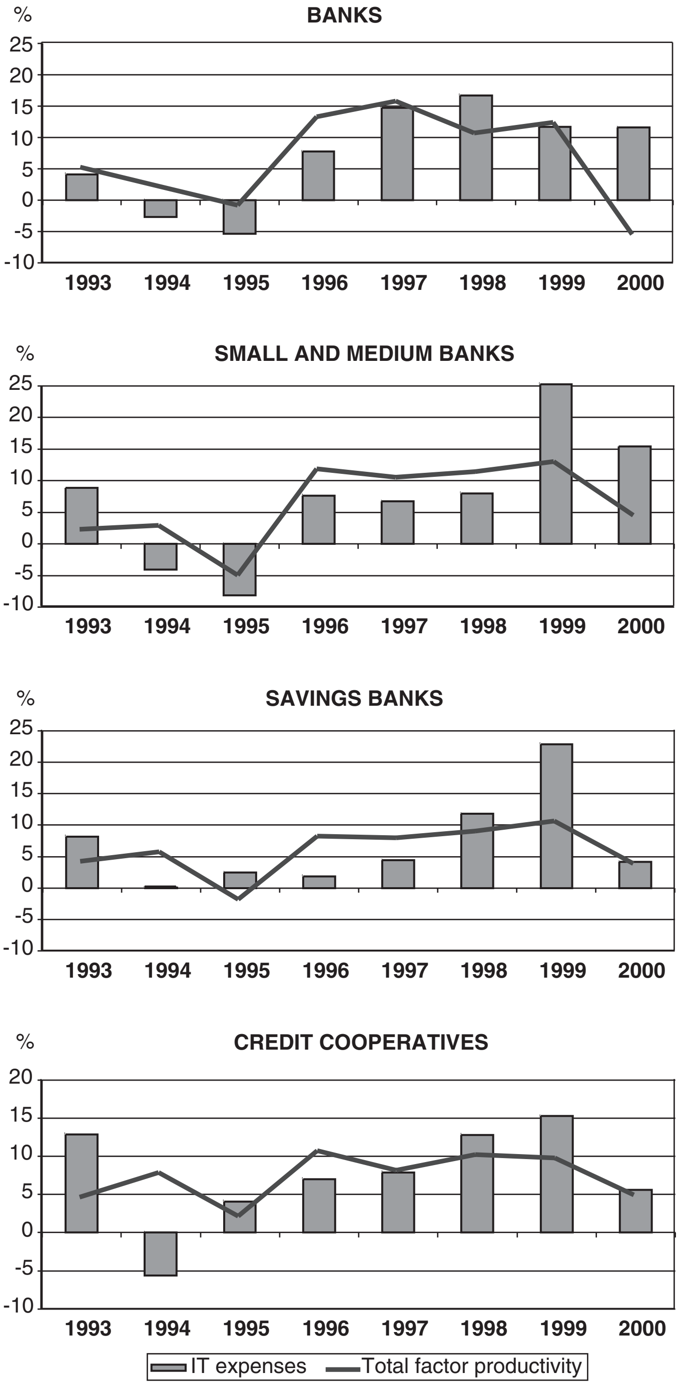


Figure 5.5 Total factor productivity (rate of change).

II Risk control and new technology

The application of new technologies in banks' production and distribution processes is expected to improve bank productivity and efficiency but, at the same time, it also leads to changes in the level and structure of the risks faced by banks. The introduction of new techniques in the production and distribution of banking services does not involve any change in the categories and definitions of the risks traditionally associated with banking activity. Yet it does introduce certain changes in the way such risks are generated, as well as opening up new channels through which banks can be exposed to risks that might have a negative impact on their net worth.

The new management systems established by banks are helping to improve the overall management of risk in their activity but they also give rise to other elements of risk. These arise from inadequate application of such systems, from their possible failures and from insufficient knowledge on the part of managers and/or staff of the banks as to how they operate. Meanwhile, the application of new product distribution techniques, in particular those using the Internet, is also a source of potential risks for banks. Moreover, the increasing use by banks of outsourcing for certain services, whilst helping cut operating costs, may also become a potential source of risk. The decision to invest in new technologies, which are often very expensive and potentially have a very short useful life, may be a potential source of risk too. Furthermore, the very decisions on the orientation of bank business arising from the impact thereon of the emergence of these new technologies, are another source of potential risks for banks. Finally, the application of new technologies and their impact on certain areas of business may help introduce changes to the risks arising from traditional banking activities.

There are therefore several channels through which the application of new technologies[5] may affect banks' risk structure, which means that it is necessary for both the banks and the supervisory authorities to adapt their current models for risk control and management to the new situation. The speed and intensity of the change in the risk profile of banks will depend on the degree of application of these new techniques as well as on the vulnerability of the banking sector of each country to competition from other financial service providers.

In the case of Spain, change may occur rapidly, given that the main banks have made considerable efforts to renew their management, production and distribution systems in order to harness the advantages arising from the new technologies. Moreover, the Spanish banking model based on retail business, with an extensive branch network to distribute products, may be very sensitive to changes in customer preferences regarding the preferred way of carrying out banking transactions. Finally, the high level of competition in the Spanish banking sector has meant that the banks have made significant efforts to make the most advanced systems of bank product distribution available to their customers, developing Internet banking divisions or setting up subsidiaries to engage specifically in such business.

Table 5.3 Internet and e-banking penetration

Country	*% Internet penetration (% of population)*	*% e-banking penetration (% of population)*
Finland	44.0	37.6
Sweden	55.3	27.9
Norway	52.8	27.8
Denmark	48.3	14.7
Switzerland	33.1	7.9
United Kingdom	33.1	7.8
Holland	45.5	6.4
Spain	13.6	5.0
United States	55.9	4.1
Germany	24.3	3.6
France	15.6	2.4
Italy	23.3	2.0

Source: OECD, FPK and www.nua.com.

These strategies by Spanish banks have led to more rapid development of electronic banking in Spain than in other countries where Internet use is higher (see Table 5.3). Spain is among those European countries having a higher number of electronic banking customers as a proportion of population, although many of the customer transactions made through electronic banking media are probably merely requests for information on products or prices and do not entail the contracting of specific financial services. The market share of Internet banks in the business of all deposit-money institutions is still very low, reaching a significant amount only in the case of current accounts where it exceeds two per cent of the total and in savings accounts with almost 0.9 per cent. However, these figures only refer to the contracting of financial services through banks operating solely through the Internet and do not include the contracting of services through the Internet divisions of traditional banks which may be more important.

The scant importance of Internet business for the moment is expected to increase at a good pace, since the growth rates recorded by the group of electronic banks is very high, especially in recent months. Yet it seems that some time will be needed before electronic banks achieve significant market shares in the Spanish banking industry.

Categories of risk

Although, as already mentioned, the categories of risk associated with the use of new technologies are not new, they do have certain characteristics which make it necessary to revise current risk control mechanisms.

Based on consultations of the supervisory authorities in various countries, a classification has been established of the risks associated with new technologies which would need to be taken into account when evaluating the situation of the banks which make up the banking system of each country. These categories are as follows:

- *Strategic and business risk* – arising from the consequences that strategic management decisions relating to the challenges posed by technological development may have for the evolution of the bank.
- *Operational risk* – emerging as a consequence of the growing use of new technologies in the banking industry, leading to greater risks associated with improper use or operation of the new systems applied.
- *Reputation risk* – the possibility of losses arising from improper operation of the new systems helping to undermine the bank's good image with its customers.
- *Legal risk* – risks arising from breach of regulations or negligence that may entail the payment of compensation or fines.
- *Other risks* – a category that basically includes the rest of the risks traditionally associated with banking activity (credit, exchange rate, liquidity, market, systemic) which are also influenced by the introduction of new technologies.

The changes that the adoption of IT technology may give rise to in each of these risk categories shall now be analysed in greater detail, with particular reference to the Spanish banking system. Though the process of change in risk profile is common to all countries with a developed banking system, it has certain specific characteristics according to the structure of business of the banks that make up the system.

Strategic and business risk

This is the broadest of the categories considered and the decisions taken in this area will have effects on the other categories considered. The revolution in data processing and transmission technologies, together with other elements of change,[5] has entailed a radical transformation in the banks' operating environment. This has increased the uncertainty over the appropriate strategy, raising the risk of losses arising from wrong decisions.

The new technologies increase the level of competition in banking markets,[6] open up new possibilities for expansion into other markets, make possible potentially cost-saving changes in the structure of production, improve the efficiency of mechanisms for data management, product design and risk control and introduce a radical change in product distribution systems.

All of these changes require strategic responses from bank management in order to put the bank in a position from which it is able to compete in this new situation. In the case of Spain, and given its traditional business structure centred on retail banking with a very dense branch network, the possibility of developing alternative distribution networks based on Internet use, with much lower costs than traditional mechanisms,[7] may involve a very significant decline in the strategic value of the branch network, creating a serious problem of overcapacity for the banks.

However, the development of these alternative distribution networks is not taking place as fast as expected. Despite the advertising drives by institutions more oriented to this business segment, the results are still not very brilliant. The physical presence of the bank with which they operate still dominates the preferences of Spanish bank customers. Also, the expected competition from abroad and from other non-financial institutions is still not particularly important and the contracting of financial services through the Internet is being hampered by the uncertainty regarding the security of transactions and by the legal difficulties involved in formalizing contracts through electronic media.

All this means that the impact of electronic banking on the traditional banks is still very limited. Even so, it should be noted that some of the problems of security and legal difficulties will undoubtedly be resolved in the near future and that the growth rates of banks operating exclusively through the Internet have, in recent months, been very much higher than those for deposit-money institutions in general. Accordingly, banks will have to develop strategies aiming to safeguard their position in the market, namely plans to reduce operating costs, to develop alternative distribution channels and to eliminate possible overcapacity in the traditional distribution network.

However, the possible defensive strategies of traditional banks are not risk-free. A policy of expanding the customer base through aggressive pricing could compromise the future profitability of these banks, exacerbating even more the possible problems arising from the overcapacity of their networks. In this respect, the aggressive strategy to attract customers used by some electronic banking institutions, offering above market interest rates, could tempt traditional banks to adopt a similar strategy, with the consequent risk to their financial stability.

The policies of some electronic banking institutions to attract customers, based on price reductions, may be having a 'cannibalization' effect similar to that seen with the appearance of high interest accounts at the end of the 1980s. This is because a significant proportion of the new customers come from the customer base of the traditional banks which are the parent companies of these electronic banks. This may hamper the achievement of foreseen targets when investing in new distribution channel technologies. In that case, these policies would not help widen the customer base of the traditional business but merely shift customers from the traditional banks towards the electronic institutions, which have narrower operating margins. Despite these possible negative effects, there is perhaps no better business alternative since failure to introduce these new systems may be an even worse option owing to the loss of those customers who contribute higher value added to the bank's business.

Lastly, there is another strategic risk associated with technological innovation which is related to decisions to invest in technology (both for internal use and to develop external distribution systems) and also to agreements with non-financial companies (particularly in the telecommunications sector) to develop joint systems to provide financial services using the latest technological advances. The investment required to establish such systems is very large and its useful life, owing to rapid technological innovation, may be very short, so that decisions in this area must be carefully considered in order to avoid wasting resources on investments of doubtful profitability. However, at the same time, strategic decisions in this area of business cannot be put off without running the danger of being sidelined from new markets. The fact that the managers of deposit-money institutions have little knowledge of the characteristics of these technologies may entail an additional source of concern.

In the case of Spain, the existence of a large number of small institutions, which find it more difficult to raise funds and have a lower technical capacity to acquire an adequate level of knowledge in this area, might suggest that the risk of taking wrong decisions is greater. Nonetheless, the existence of associations both in the case of the savings banks and the credit cooperatives, helps to mitigate such risk, since these can be used to address this kind of strategic decision, developing systems for joint use and thereby making available more resources and qualified personnel.

Operational risk

The growing use of new information processing and transmission technologies enables banks to exploit their databases more effectively, to generate new business opportunities and to develop risk control models which help to improve management. At the same time, the possibility of setting up new distribution channels as well as of developing financial products that are better adapted to customer requirements is also a new source of income and increases the chances of establishing stable relationships with customers which help to increase their loyalty. Moreover, the possibility of restructuring productive processes by subcontracting some non-essential services may help to reduce operating expenses. Together with these beneficial effects, new technology entails some costs for banking institutions, since it makes them more vulnerable to possible failures in their operating systems. This is what is known as operational risk.

The development of all kinds of systems based on the intensive use of technology increases the exposure of banks to failures in their operation, to their fraudulent use by third persons or employees and to failures in external systems that they use to supplement their own. At the same time, the growing complexity of the systems used makes it necessary to train employees and management, to ensure that they have adequate knowledge of how these tools work. The growing importance of this kind of risk has led supervisory authorities to introduce in the new Basel Capital Accord specific capital requirements to cover the risks arising from operating system failures.

In the Spanish banking market this kind of risk may be rather significant owing to the complexity of the banking groups. The universal banking system predominates, with financial conglomerates offering all kinds of products and services with very diverse characteristics. This makes it more difficult to develop an integrated system for distribution and for risk management. Moreover, this multiproduct business structure, with ramifications in different markets and with different subsystems operating in parallel, tends to increase the risks that may arise from operating system failures.

Therefore, it is particularly necessary for the banks to develop rigorous control systems for these new applications so that they can control any failure in their risk management models, in the transmission of data from one business unit to another, or any weakness in security and database information access systems that allowed fraudulent access by employees or third persons. Also, banks should establish systems to ensure control and monitoring of the quality and security of firms subcontracted in their productive processes.

Owing to the complexity of Spanish banking groups, which means that the control of their activity is divided up between different supervisory authorities, an additional effort is required by the latter to ensure that effective control exists over all areas of business and all the internal and external data-processing systems involved in the correct operation of banks.

Reputation risk

Closely related to the previous point, the growing complexity of banks' data-processing systems, with a higher degree of involvement of external systems over which there is not always adequate control (subcontracted services, alliances with non-financial firms and development of distribution systems based on the Internet or other external networks), increases the possibility of failure in the correct provision of services. In a relationship based on confidence in the proper working of the bank, where the decision to deal with a particular bank has a lot to do with the prestige of the brand, any failure in systems that casts doubt on the reliability or security of transactions may have a very negative effect on customers, leading to a flight to other competing banks. The risk of a loss of 'prestige' among customers may eventually lead to a loss of market share that is difficult to recover in such a competitive environment as that expected in the years ahead.

This is another of the most prevalent risks in the Spanish banking system, which is based on a close relationship between the customer and the bank, the reputation of the brand being a significant factor for maintaining a stable customer base.

Legal risk

Legal risk is also closely related to the growing complexity of banks' data distribution and storage systems as a consequence of the application of new technologies. By increasing the possibilities for fraud or improper use of the information contained in bank databases, as well as the possibilities for malfunctioning of systems which damages the quality of service, the risk also increases of claims against the bank by customers, who owing to growing competition are also becoming more demanding.

In principle, the use of new distribution channels enables business to be expanded into an unlimited number of markets and countries with a very small additional cost, thus increasing the potential customer base. However, it has to be taken into account that an expansion of this type may give rise to an increase in legal risk by increasing the probability of infringing the laws in force in other countries owing to lack of knowledge thereof. Although the first steps have already been taken to try to adopt a more or less uniform legal framework to facilitate the cross-border marketing of financial products through the Internet, the current situation tends to apply to such transactions the laws of the country in which the products are offered. This tends to increase the possibility of infringing the law applicable, with the consequent legal risk involved.

For the moment, the provision of cross-border banking services through the Internet is very limited and, in almost all cases, there is already a physical branch in the country in which it is carried out, which helps minimize the risks of failing to comply with regulations.

Other risks

The application of the new technologies has an effect on the traditional risks of banking activity which is difficult to calculate. On the one hand the new techniques enable these risks to be better managed, which helps reduce exposure. On the other hand, however, certain policies associated with the changes induced by the appearance of these new technologies may tend to increase some of these risks.

In the case of Spain, the clearest effect would be a possible increase in credit risk if the banks adopt aggressive expansion policies as a consequence of the increase in competition caused by the new distribution channels. Another possible negative effect would be an increase in liquidity risk as a consequence of greater volatility of customer deposits. The possibility that the Internet gives of comparing prices and choosing the best offer may tend to weaken relationships between customers and their bank making them more sensitive to prices than to other considerations.

Since the introduction of risk management models has a positive effect on banks' levels of credit, interest rate and market risk, there is also a risk of excessive confidence in them which may give rise to problems at times of crisis. Finally, the subcontracting of services could lead to an excessive concentration in certain systems or operators, leading to an increase in systemic risk in the event of failure of such operators. The fact that some of them, owing to their status as nonfinancial institutions, may not be subject to the control of any supervisory authority makes more serious this possibility.

Regarding the attitude of supervisory authorities, being aware of all the already mentioned potential risks for banking institutions, they have started to work in the design of new systems for control. In the case of Spanish authorities, a group has been set up within the Directorate General of Supervision of Banco de España to analyse the risks arising from electronic banking.

III Conclusions

The introduction of new technologies in the banking industry involves, in general, an increase in its efficiency and competitiveness, which will probably result in benefits for customers in the form of lower prices and a better quality of service. In the case of Spanish banking institutions, there appears to be an improvement in some measures of efficiency and productivity which seem to be related to investment in information and communication technologies. At the same time, there are signs of an increase in the average qualification of workers and, correspondingly, in their wages. However, the nature of the data does not allow appropriate testing of these hypotheses due to the small number of observations and their aggregated level, which may obscure possible regular patterns if one or two big institutions deviate from those regularities. This evidences the necessity to use data from individual banks to avoid these biases.

Moreover, the application of new systems for risk management and for processing and transmitting data is expected to help improve the soundness of banks. However, during the transition period, a number of uncertainties may arise making greater control by the supervisory authorities necessary in order to ensure the solidity of banks.

The application of new technologies to the banking services industry entails a wide range of changes in the risk structure of banks. This warrants a reform of the current risk control systems in order to adapt them to the new conditions of the sector. To avoid any delay in the introduction of these new risk control systems it is necessary to start working on them in the current take-off phase of the new banking services. Being aware of this the various supervisory authorities have already been designing a control system for some years.[8] As regards the Spanish authorities, a specific group has been set up in Banco de España to analyse the risks arising from electronic banking. This group has already started the necessary work to develop a system to assess, monitor and control this kind of risk in the Spanish banking system.[9]

Notes

1 CECA (Spanish Confederation of Savings Banks) is an association which performs tasks of representation of all the savings banks and provides advisory and securities management services to many of them. Banco Cooperativo is an institution in which a very high proportion of credit cooperatives have holdings. It performs a similar role to that of CECA in the savings bank sector.

2 The decline in the average cost per transaction seen in small and medium-sized commercial banks in the last two years does not reflect a genuine cost reduction, but is related to changes in tax rules which stimulated the issuance and placement of short-term bank securities (commercial paper) among customers, sharply increasing the number of transactions performed.

3 If hours worked are used instead of the number of employees, the change in productivity is very similar, since the rate of growth of both variables displays a very similar time profile.
4 See Pérez and Pastor (1994) and Más and Pérez (1990) on the equivalence of this measure, called *revealed productivity*, and the total factor productivity in Solow's analysis.
5 Many of these changes affecting banking activity are influenced by other factors such as the general process of globalization of financial activity, the increasing economic integration between different countries and ongoing deregulation of the financial system.
6 See Vesala (2000).
7 According to some reports on the US market, the cost per transaction through branches is almost 100 times greater than the cost through the Internet.
8 See the various documents of the Electronic Banking Group of the Basel Committee and the new BIS II Capital Accord.
9 The views expressed in this paper are the authors' and do not necessarily reflect those of the Banco de España

References

Analistas Financieros Internacionales (2000) 'El impacto de la banca on-line', AFI nota técnica no. 18.

Basel Committee on Banking Supervision (2000) 'Electronic Banking Group Initiatives and White Papers', BIS.

Basel Committee on Banking Supervision (2001) 'Risk management principles for electronic banking', BIS.

Lascelles, D. (2000) 'Regulating financial services on the Internet', *Central Banking* 10(2).

Más, M. and Pérez, F. (1990) 'Productividad revelada: un análisis de costes', *Investigaciones Económicas*, suplemento 71–6.

OECD (2001) 'Cross-border trade in financial services – implications of electronic finance', DAFFE/CMC (2001) 12.

Pérez, F. and Pastor, J.M. (1994) 'La productividad del sistema bancario Español (1986–1992)', *Papeles de Economía Española* no. 58, 62–86.

Vesala, J. (2000) 'Technological transformation and retail banking competition: implications and measurement', *Bank of Finland Studies* E–20.

6 Consumer behaviour and the usage and adoption of remote and direct banking in the United Kingdom*

Barry Howcroft

Abstract

Against a background of change in society, and high levels of competition and developments in remote delivery channels within the financial services markets, this paper attempts to ascertain what effects these changes have had, or are likely to have, on bank customer behaviour. This paper outlines a conceptual model of consumer behaviour and subsequently tests it by applying cluster analysis to a dataset based on a questionnaire of bank customers. The results suggest that the traditional perception of the bank customer as essentially inert and passive might no longer apply. The paper accordingly discusses the essential characteristics of the emerging customer segments and assesses some of the strategic implications for banks.

1 Introduction

The market for financial services has become increasingly competitive with unprecedented levels of new entrants and developments in remote delivery channels. Society in general has also seen radical changes with, amongst other things, a noticeable trend towards greater consumer empowerment. For example, in the United Kingdom individuals are being actively encouraged by the Government to take personal responsibility for their pensions and tax returns. Against this background of change both within the financial services markets and society, the paper attempts to ascertain what effects these trends have had, or are likely to have, on bank customer behaviour.

The traditional perception of the bank customer was perhaps best captured by Brown (1952) when he coined the term ‘behavioural loyalty’. When applied to bank customers it suggests that although the majority of bank customers do not regularly change banks, their reluctance to change is not borne out of any strong feelings of brand loyalty. Rather loyalty in this context is derived more from feelings of inertia or apathy and the high time and transaction costs associated with changing banks. Underpinning behavioural loyalty, therefore, is the fundamental belief that there is little point in changing banks because there is no discernible difference between them in terms of product offerings, price, service quality, etc.

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

One possible way of ameliorating this problem is to focus upon the concept of 'customer value' (Woodruff, 1997) which involves financial providers adopting an external orientation and seeking to understand more fully the nature of their customers. A critical feature of this approach is to initially ascertain and then subsequently influence and determine consumers' buying behaviour. Accordingly this paper outlines a conceptual model for ascertaining changes in consumer behaviour within the context of financial services. This model is subsequently tested by a questionnaire and appraised through the use of cluster analysis procedures. Section 2 of this paper provides a review of the relevant literature and develops the conceptual model. Section 3 outlines the research methodology, Section 4 assesses the findings from the study and Section 5 draws some conclusions.

2 Understanding consumer behaviour and financial services

The need to further academic understanding of consumer behaviour within the financial services sector has become a critical issue for a number of reasons, not least being the impact of deregulation and the emergence of new forms of technology. At the same time academics have suggested that a conceptual framework is currently lacking which adequately explains how consumers purchase financial services (McKechnie, 1992; Betts, 1994).

An important step in understanding how consumers purchase financial services is to ascertain how they interact with financial service providers. A review of the literature on bank consumer–buyer interactions is revealing insomuch as price does not feature very prominently. This is probably due to the fact that within the UK certain bank services, especially current accounts and those associated with transaction banking, have been provided free of charge (if in credit) to consumers. Moreover, when charges are levied they tend to be so difficult to understand that consumers find it almost impossible to make a rational choice based on conventional price–quality criteria (Drake and Llewellyn, 1995).

Accordingly, the literature on consumer–buyer interactions (Bateson, 1989; McKechnie, 1992; Harrison, 1994; Ennew and McKechnie, 1998) suggests that two factors principally motivate and determine individual purchasing or contracting choices, namely *involvement* and *uncertainty*. Consumer 'involvement' in the buyer–seller interchange incorporates a number of subsets: customer control (Bateson, 1989), customer participation, and level of contact (Chase, 1978). Chase (1978), for example, offered a framework to distinguish between services based upon level of contact during the service delivery stage. Consumer uncertainty has been analysed in terms of consumers' perceptions of risk and research suggests that this is ultimately related to the characteristics of the product (Shostack, 1977; Ennew and McKechnie, 1998). Such characteristics directly influence consumer uncertainty by determining perceptions of complexity associated with different financial products (Harrison, 1994). With the development of remote and direct delivery channels, i.e. multimedia kiosks, PC banking, the Internet, telephone banking, etc., the product and the delivery channel have become inextricably linked. This suggests that the consumers' perceptions of risk and uncertainty are also related to the 'type' of delivery channel being used or on offer.

By placing consumer involvement and consumer uncertainty onto a simple continuum running from high to low, it is possible to construct a two-dimensional matrix of consumer behaviour (see Figure 6.1). This matrix describes the purchasing/contracting alternatives available to financial service consumers and, as such, reflects a range of interaction modes between financial service providers and their customers. Each quadrant represents a different combination of involvement and uncertainty (referred to as consumer confidence in the matrix) and, therefore, different modes of banker–customer interactions when purchasing financial products.

These considerations are significant because the limitation of previous consumer behaviour models (Nicosia, 1966; Engel, Kollat and Blackwell, 1968; Howard and Sheth, 1969; Bettman, 1979) stems from the factors hypothesized to explain consumer motivation and behaviour, particularly in the creation and maintenance of relationships. Previous models have tended to focus on the extensive role of information in shaping and directing rational choice. Nicosia (1966), for example, identifies the search and evaluation of information; Howard and Sheth (1969) link the amount of information held with the extent of learning and thus problem solving; Engel, Kollat and Blackwell (1968) identify information input and processing as central elements in decision-making; Bettman (1979) recognizes that individuals have limited capabilities for processing information and this limitation affects their decision-making activities. The common causal link in all of these approaches can be summarized as: information – attitude – purchase. As an explanation of behaviour this form of linkage is not only predominant in the social sciences, but can also be clearly identified in behavioural economics (Katona, 1960; Scitovsky, 1976). All of these models are based on information which is necessarily assumed to be freely and readily available with little reference made to the context in which information is found and used. Academics have thus argued that

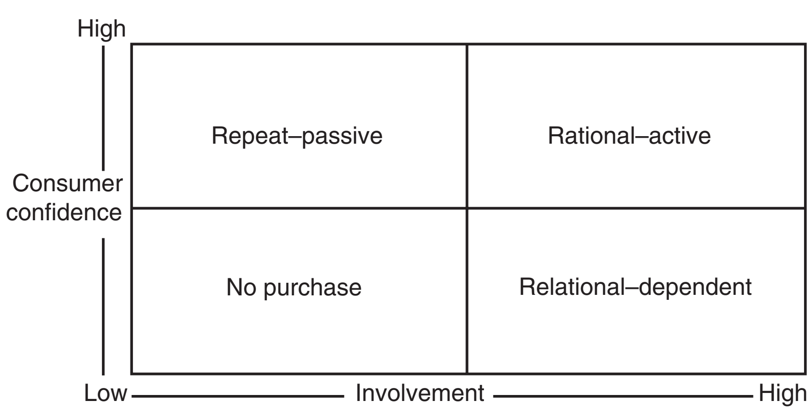


Figure 6.1 Consumer behaviour matrix.

such models are largely unverifiable empirically because they do not offer testable hypotheses (Tuck, 1976; Foxall, 1991). These earlier models also ignored consumer involvement with the purchase and the consumer's ability or confidence when making purchase decisions (Morgan and Hunt, 1994). Their specific applicability to the field of financial services can also be questioned because they typically refer to one-off purchases rather than the recurrent and long-term relations common in financial services (McKechnie, 1992).

In contrast, the proposed consumer behaviour matrix (see Figure 6.1) is based upon theoretical insights from the work of Dwyer, Schurr and Oh (1987) and Thibaut and Kelly (1959). Methodologically the approach seeks to utilize the notion of the 'ideal type' promulgated by Weber (1949). Weber sought to characterize phenomena into broad groups or ideal types which represented the self-conscious and 'one-sided accentuation' of the most significant features of social phenomena. Once classified into groups, the constituent elements of the phenomena being observed are available for analysis. The rationale for this approach is that complex social interactions rarely, if ever, operate to a set pattern like elements in the physical sciences. Rather than postulate general theories that describe behaviour in all contexts, ideal types describe forms of behaviour in certain contexts, which facilitate the construction of hypotheses. The ideal type expresses what Weber described as an 'objectively possible' course of actions within a relevant frame of reference (Parsons, 1951). Behaviour is, therefore, dependent upon both the nature and the context in which it occurs. An ideal type will consequently possess a number of distinguishing characteristics, constructed by the researcher to describe the complexities of behaviour within a particular frame of reference.

Weber argued that the key to developing ideal types was to identify characteristics that shaped both the forms of the ideal type and the frames of reference. It follows that in developing ideal types which characterize consumer buying, it is necessary to identify the underlying constructs which determine consumer behaviour within a particular environment or frame of reference. Ideal types, therefore, reflect certain underlying constructs or imperatives, and individual buying behaviour is determined accordingly. This is consistent with the work of Fishbein (1967) which links attitudes and outcomes, arguing that individuals' attitudes toward certain outcomes motivate behaviour.

The advantage of the consumer behaviour matrix, as an approach to understanding consumer buying and contracting behaviour, is that it draws upon a diverse source of literature, including economics (Simon, 1957), consumer behaviour (Bloch, 1982; Bloch and Richins, 1983) and psychology (Thibaut and Kelly, 1959), and combines these within a single framework. The rich diversity of literature in these areas is brought together to create ideal types of behaviour which can be applied to actual buying and contracting in financial services. In this way, the consumer behaviour matrix places observed behaviour within a context. By identifying the underlying factors influencing buying and contracting behaviour and linking these directly with consumer needs, the rationale for consumer interactions and relationships can be more readily analysed.

As such the consumer behaviour matrix recognizes that financial services are not homogeneous but rather can be regarded as being on a continuum ranging from the relatively straightforward and easy to understand at the one extreme, through to the highly complicated and difficult to understand at the other (Shostack, 1977). Similarly, the matrix also recognizes that consumers might be quite markedly different in terms of their levels of financial literacy and in their abilities to make financial decisions about whether or not to make purchases of financial services.

Furthermore, the use of this framework permits the adoption of a deductive approach in the choice of a clustering variate. The technique thereby facilitates segmentation on a range of self-selected dimensions (Christopher, 1969). As Ketchen and Shook (1996, p. 451) highlight: 'Cluster analysis has played a key role in … [strategic management research] because it allows for the inclusion of multiple variables as sources of configuration definition, thus enabling the creation of potentially rich descriptions.' Harrison (1998) has adopted such a cluster-based approach in analysing the financial services consumption patterns of UK bank customers. Using insights from qualitative research (Harrison, 1994), an attempt was made to replicate and validate the analysis through cluster analysis. The research suggests that consumers differ in terms of their perceived knowledge and understanding of financial services. They also differ in their confidence and ability to handle financial matters and in their general interest in the subject matter. On these bases, Harrison drew attention to four distinct consumer clusters, ranging from the 'financially confused' who were characterized by low levels of both perceived knowledge and financial maturity to 'capital accumulators' who were characterized by high levels of perceived knowledge and a high degree of financial maturity. McDougall and Levesque (1994) in a similar fashion sought to segment consumers on the basis of service quality. Their research was able to draw attention to two segments: those consumers who were primarily convenience-orientated and those who placed priority upon performance characteristics.

In an endeavour to make the matrix reflect consumer behaviour when purchasing a fairly comprehensive range of heterogeneous financial services it was important to review the literature as the classification of financial services. Services can be classified in a number of differing ways (Lovelock, 1983; Normann and Haikola, 1986; Storbacka, 1994). They can be grouped in terms of their marketing characteristics, such as the nature of the provider–customer relationship (Lovelock, 1983) or on the basis of the content of the interaction between provider and customer (Storbacka, 1994). Storbacka proposed a typology of interactions distinguishable in terms of the duration and frequency of the interaction, and the level of customer control, contact and participation. Using this typology he drew attention to five main governance systems: transactions, deposit and lending, counselling, specialist services and investment services. This paper utilizes the Storbacka typology, but excludes counselling on the basis that it is implicit in the other forms of governance. Four categories of financial services are, therefore, highlighted as representative of the range of interactional contexts: transactions-based services such as the 'primary' current account, i.e. where consumers have more than one current account, the one that is most frequently used; insurance-based services such as house contents, buildings and motor insurance; credit-based services such as personal loans and mortgages, and finally, investment-based services which are perceived as being relatively complicated with commensurately high levels of risk, i.e. PEPs, TESSAs, stocks and shares and pensions. In this manner the research sought to segment consumer behaviour on the two constructs of involvement and uncertainty for a range of differing financial products.

Table 6.1 shows that the four categories of financial services were fairly well represented across the sample. However, it is interesting to note that although 39.5 per cent of the respondents were connected to the Internet, only one respondent had used the Internet to purchase a financial product.

Table 6.2 reveals that the market is changing insomuch as when questioned about their preferred acquisition channels, the respondents revealed a noticeable shift towards acquisition via remote and direct delivery channels, i.e. ATMs, the telephone and the Internet. In particular, there is a distinct move towards telephone banking and a definite, albeit modest, move towards acquisition over the Internet. Table 6.2 also shows that respondents had a preference for using more than one delivery channel per product.

2.1 Forms of consumer behaviour

The model outlines four ideal types of consumer behaviour which are labelled: repeat–passive, rational–active, no purchase and relational–dependent. These quadrants have previously been examined in the light of qualitative research (Beckett, Hewer and Howcroft, 2000); however, a brief description of the chief characteristics of each quadrant follows.

Repeat–passive

Consumers display low levels of involvement with the financial product as they are fully aware of the product's salient features. Given the low levels of involvement

Table 6.1 Consumers' acquisition of financial product by type of delivery channel

	Current account		*Insurance-based*		*Credit-based*		*Investment-based*	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
At the branch	202	85.6	53	21.7	124	70	83	44.2
Visit at my home	5	2.1	28	11.5	13	7.3	44	23.4
Over the telephone	17	7.2	134	54.9	23	13	13	6.9
Over the Internet	0	0	0	0	0	0	1	0.5
By telephone and post	12	5.1	29	11.9	17	9.7	48	25.5
Responses	236	100	244	100	177	100	188	100

N = 244

Table 6.2 Consumers' preferred acquisition channel

	Current account		*Insurance-based*		*Credit-based*		*Investment-based*	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
At the branch/office	204	68.2	107	38.6	164	62.8	171	65.2
At an ATM	16	5.3	0	0	3	1.1	0	0
Over the telephone	64	21.4	150	54.2	79	30.3	73	27.9
Over the Internet	15	5.1	20	7.2	15	5.8	18	6.9
Responses	299	100	277	100	261	100	262	100

Chi-square: value 112.135; df 9; Asymp. Sig. (two-sided) .000
No. of valid responses: 1099

and the limited perception of uncertainty, these consumers can be described as 'passive' in the sense that they will make repeated interactions without actively seeking alternatives. This repeated pattern of purchase behaviour, which is described as 'behavioural loyalty' in the literature, has been extensively researched (Brown, 1952; Johnson, 1973, 1982).

Rational–active

These consumers display high levels of involvement in terms of the process dimensions of control, participation and contact. Confidence in terms of understanding the product and certainty of *ex post* outcome is similarly high. To purchase in an 'instrumentally rational' manner, individual consumers are assumed to possess sufficient ability and information to enable them to make clear comparisons between competing products and thus make an informed choice. If the information is not available or the consumer lacks the ability to make choices, they have to move away from 'instrumental' rationality as discrete contracting is no longer an effective means of structuring the transactions.

No purchase

This quadrant describes consumers who, because they have no involvement with a financial provider and do not possess the ability or the confidence to make transaction decisions, make no purchase.

Relational–dependent

In this quadrant consumers are highly involved, but are not very confident due to the complexity of the product and uncertainty of *ex post* outcomes. In order to make choices, the consumer will seek advice and help from banks or third parties and can, therefore, be described as 'dependent consumers' who form relationships to reduce uncertainty and structure their pattern of purchases.

Relational contracting does not fit easily into the concept of either an active or a passive interaction, but it is clearly an important aspect of the banker–customer relationship. It emerged from the work of MacNeil (1978) and Williamson (1975, 1985), who recognized that in particular contexts rational–active and repeat–passive contracting were not effective in structuring exchange. It is typically used in highly uncertain environments where consumers lack the information to make rational decisions, but, nevertheless, perceive differences in quality between competing products or services. Under these circumstances, if consumers want to make informed choices, they have to draw on the assistance of more informed third parties. This relationship with third parties effectively replaces the information search and processing activities found in repeat–passive and rational–active contracting. Trust plays a critical role in this relationship and, to a large extent, the role of professional associations is to protect consumers from the opportunistic behaviour of third parties.

3 Research analysis and sample

The previous section provided a detailed examination of the relevant literature relating to consumer behaviour in the context of purchasing/contracting financial products and a conceptual model outlining four ideal types of consumer behaviour (see Figure 6.1) was proposed. The main objective of this research was to examine consumer behaviour and ascertain if the conceptual model could be supported empirically. In addition, the research wanted to ascertain whether the majority of bank consumers in the sample conformed to the traditional view of the bank customer as being behaviourally loyal, as represented by the repeat–passive quadrant in Figure 6.1. The paper also attempts to profile each of the quadrants and any other significant clusters revealed by the analysis, provide an explanation as to why consumer behaviour might be changing and identify some important strategic implications of these changes for banks.

Following a series of focus discussion groups, a comprehensive questionnaire was designed and piloted using university personnel. The definitive questionnaire was then sent to 4000 UK consumers. This resulted in 244 usable responses, i.e. out of the 351 responses, 244 respondents had answered all of the questions. The relatively low response rate of fewer than ten per cent was, with retrospect, not that surprising given its length and the confidential nature of the information requested. It did, however, raise a number of important issues, namely, to what extent was the low response rate also indicative of the respondents' disinterest in financial affairs. This is important because the research was attempting, amongst other things, to ascertain the extent to which bank customers can still be described as behaviourally loyal. In this respect the non-respondents might be indicative of a large section of bank customers who can best be described as inert or passive and, therefore, conform to the definition of a behaviourally loyal customer. Another major concern was the representativeness of the sample. In this respect, when compared with recent UK statistics (Social Trends, 1999; Advertising Association, 1998), it was found that the sample was over-representative of middle age groups (35–45-year-olds) and under-representative of younger consumers (between 18 and 24). The sample was also biased towards consumers in the higher income groups (i.e. over £30,000). This profile might possibly be explained by the fact that wealthier consumers have a tendency to be older and, therefore, have relatively more experience and interest in financial services than their younger counterparts.

This paper uses only part of the original questionnaire but the questions and variables used in the analysis are shown in Table 6.3. Variables 1–20 inclusive focus on the two dimensions highlighted by the consumer behaviour matrix (Figure 6.1): consumer confidence and consumer involvement. Specifically, variables 1–4 and 5–8 relate to consumer deliberation and the importance of talking to somebody. As such, they were interpreted as indicators of consumer

Table 6.3 Questionnaire items used for cluster variate

Q. no.	*Question*	*Var. no.*
	Financial services attitudes:	
Q1	I tend to deliberate a lot when changing…	
	my current account	V1
	my insurance products	V2
	my credit-based products	V3
	my investment-based products	V4
Q2	It is important to talk to somebody when selecting…	
	my current account	V5
	my insurance products	V6
	my credit-based products	V7
	my investment-based products	V8
Q3	I possess a good knowledge and understanding of...	
	my current account	V9
	my insurance products	V10
	my credit-based products	V11
	my investment-based products	V12
Q4	I felt unsure when selecting…	
	my current account	V13
	my insurance products	V14
	my credit-based products	V15
	my investment-based products	V16
Q5	Over the past two years I have considered switching…	
	my current account	V17
	my insurance products	V18
	my credit-based products	V19
	my investment-based products	V20
	Socio-demographic variables:	
Q6	What is your current (gross) annual household income?	V21
Q7	Age last birthday	V22
Q8	Which of the following qualifications do you possess?	V23
Q9	How many dependent children do you have?	V24

involvement. A high emphasis on these variables was interpreted as being indicative of either rational–active or relational–dependent behaviour and it was regarded as being equally conducive to the use of branch networks, home visits or the telephone but not Internet banking. Conversely a low emphasis on consumer involvement was postulated as being indicative of either repeat–passive or no-purchase behaviour. Variables 9–12 and 13–16 relate to consumer knowledge and understanding, and certainty when selecting financial products. As such, they were taken as indicators of consumer confidence. High levels of confidence were regarded as indicative of either repeat–passive or rational–active behaviour.

Having regard for the integration of delivery channels and financial products, uncertainty and a poor knowledge of financial products was interpreted as being indicative of consumers who prefer using branch networks and home visits rather than the telephone or other forms of remote banking. In terms of consumer behaviour it was also regarded as reflecting relational–dependent or no-purchase behaviour. Evidence of willingness to switch financial providers, shown by variables 17–20, was interpreted as being indicative of consumers who have a propensity to use remote or direct banking and rational–active behaviour. Conversely a reluctance to switch was associated with either repeat–passive or relational–dependent behaviour and a tendency to use branch networks. The remaining variables (21–24) provided additional insight into the socio-demographic details of the respondents.

Questions 1–5 relate to the four different categories of financial product: primary current accounts; insurance-based products; credit-based products; and investment-based products. Respondents were invited to respond on a 5-point Likert scale ranging from Strongly Disagree to Strongly Agree. Additionally, four socio-demographic questions (Q6–Q9) were also included to help to differentiate between the consumers on the basis of their income, age, education and dependent children.

3.1 Research methodology

Given that the primary aim of this research was to group consumers on the basis of their behaviour and attitude towards financial products, it was decided to use cluster analysis. This method attempts to maximize the homogeneity of objects (e.g. consumers) within the clusters, while simultaneously maximizing the heterogeneity between the clusters (Hair *et al.,* 1998, p. 470). Support for adopting this methodology comes from Punj and Stewart (1983) when they state that: '… important use of cluster analysis has been made in seeking a better understanding of buyer behaviours by identifying homogeneous groups of buyers'.

Following the approaches put forward by various writers (e.g. Hair *et al.,* 1998; Ketchen and Shook, 1996; Doherty *et al.,* 1999; Saunders, 1994), the key questions[1] when using cluster analysis would seem to be:

- How should similarity be measured?
- How should the clusters be formed?

- How many clusters/groups should be formed?
- How should the final cluster solution be validated?
- Are the clusters distinct, stable and meaningful?

These questions are addressed by outlining the research method and the results. An interpretation of the results also ensures that the final cluster solution is both meaningful and useful (Punj and Stewart, 1983).

The method used was Ward's which is both (i) a hierarchical procedure as it moves in a stepwise manner to form an entire range of clusters; and (ii) an agglomerative method as clusters are formed from the combination of existing clusters starting with each observation (case) being a cluster (Hair *et al.,* 1998). The measure of similarity used was the Squared Euclidean distance measure.

To answer the question 'How many clusters should be formed?', an examination of both the agglomeration schedule and the dendogram (Hair *et al.,* 1998) suggested that either a six-cluster solution or a five-cluster solution might be the most appropriate. Therefore, to determine the 'best' solution, from both a statistical and a practical viewpoint, a number of other factors were considered. First, both solutions were compared on the basis of distinctiveness using the Levene statistic, i.e. testing the homogeneity of variance for each variable among the clusters (Norusis, 1997), and the values of the *F* ratio, taken from the Anova. Compared with the five-cluster solution, the six-cluster solution for both measures provided more variables that were significantly different (at the 0.05 level). Further supporting evidence for the distinctiveness of the six-cluster solution came from the results of a multiple comparison procedure, the Bonferroni test (Norusis, 1998).

Second, from a practical viewpoint, although Figure 6.1 would suggest a four-cluster solution, further analysis of the characteristics of the six-cluster solution suggested that the latter solution would allow a greater understanding of consumer behaviour and would also make the model more dynamic. This issue is addressed in more detail in the interpretation section of the analysis. In order to determine whether or not the chosen six-cluster solution is stable (or reliable) as well as distinctive, a non-hierarchical clustering procedure (K-means clustering) was performed on the data using the cluster centroids from Ward's method as the initial seed points (Hair *et al.,* 1998). This approach facilitates: 'fine tuning of the results' and validation of the final cluster solution (Milligan and Cooper, 1987).

To test for an 'acceptable level of agreement' between the two clustering methods the Kappa coefficient,[2] which may be used as an objective measure of stability (Punj and Stewart, 1983), was calculated. Table 6.4 shows the consistency between the results of the two clustering methods with 192 (78.7 per cent) of the 244 cases appearing in the same clusters. The Kappa coefficient (0.743) also indicated an acceptable level of agreement (Cramer, 1998) between the two methods. Table 6.4 also highlights, in terms of the number of observations in each cluster, the similarity between the two clustering methods.

Table 6.4 Measuring agreement between Ward's method and the K-means method

		K-means method						
	Clusters	1	2	3	4	5	6	Total
	1	87.5% (28)		9.4% (3)		3.1% (1)		13.1% (32)
	2	5.4% (3)	75% (42)		8.9% (5)	3.6% (2)	7.1% (4)	23% (56)
Ward's method	3			72.1% (31)	18.6% (8)		9.3% (4)	17.6% (43)
	4	3.4% (1)	3.4% (1)	3.4% (1)	86.2% (25)	3.4% (1)		11.9% (29)
	5	2.1% (1)		6.3% (3)		81.3% (39)	10.4% (5)	19.7% (48)
	6	2.8% (1)	11.1% (4)	2.8% (1)	5.6% (2)	2.8% (1)	75% (27)	14.8% (36)
	Total	13.9% (34)	19.3% (47)	16% (39)	16.4% (40)	18% (44)	16.4% (40)	100% (244)

Measure of agreement: Kappa value 0.743.
Note: The value in parentheses is the count within each cell and total by clustering method.

3.2 Interpretation of the final cluster solution

From the statistical analysis of the final six-cluster solution it would appear that a distinctive and stable solution has been identified. However, as Doherty *et al.* (1999) argue, the real 'acid test' for the appropriateness of the final cluster solution is whether it can be meaningfully interpreted. Similarly, Saunders (1994, p. 23) suggests: 'Statistically significant results are of no consequence if they are not usable or accepted by managers.' Punj and Stewart (1983, p. 146) argue that: 'Clusters, or classes, must have demonstrable implications for hypothesis generation, theory building, prediction, or management.'

Table 6.5, as well as providing further support for the distinctiveness and reliability of the final solution, provides information that is essential to the 'interpretation' of the solution (Hair *et al.,* 1998), i.e. the final cluster centroids; the F values; and, the levels of significance for each of the 24 variables. Based upon an analysis of the F values, which test for differences between the means of the clusters (McDougall and Levesque, 1994), the most influential factors appear to be: the importance of talking to somebody for a current account (V5: $F = 45.27$) and credit-based products (V7: $F = 44.84$); knowledge and understanding for insurance and credit-based products (V10 & V11: $F = 30.17$); and uncertainty for investment (V16: $F = 25.59$) and insurance-based products (V14: $F = 25.84$). The socio-demographic characteristic which proved to be the most influential factor in differentiating between the clusters was the level of education (V23: $F = 73.67$).

Table 6.5 Six-cluster solution of the K-means cluster analysis with initial seed points from Ward's method

	Final cluster centres							
Variables	*1*	*2*	*3*	*4*	*5*	*6*	*F value*	*Sig.*
V1	2.18	3.47	2.18	3.25	3.48	2.70	11.890	.000
V2	3.26	3.70	3.51	3.95	4.27	3.45	5.866	.000
V3	3.53	3.55	3.59	3.90	4.32	3.68	3.983	.002
V4	3.79	3.68	3.95	3.85	4.36	3.68	3.067	.011
V5	3.38	4.13	1.56	2.03	3.48	2.73	45.266	.000
V6	3.26	3.55	1.56	1.90	3.34	2.85	30.800	.000
V7	4.00	4.00	2.10	2.15	4.18	3.60	44.841	.000
V8	4.00	3.91	2.49	2.90	4.05	3.83	15.573	.000
V9	4.29	4.09	4.36	4.28	4.45	3.63	6.534	.000
V10	4.18	3.79	4.10	4.08	4.25	2.63	30.173	.000
V11	4.18	3.79	4.10	4.08	4.25	2.63	30.173	.000
V12	4.18	3.32	3.69	3.73	4.02	2.58	17.146	.000
V13	1.21	2.13	1.64	1.58	1.66	1.95	7.656	.000
V14	1.24	2.60	1.90	1.75	2.00	3.10	25.842	.000
V15	1.53	2.85	2.23	1.95	2.14	3.53	23.101	.000
V16	1.41	3.15	2.56	2.40	2.18	3.65	25.593	.000
V17	1.29	2.49	3.08	2.03	2.00	2.73	9.044	.000
V18	1.21	2.83	3.44	3.18	3.73	3.10	22.512	.000
V19	1.74	2.74	3.08	2.93	2.93	2.85	6.504	.000
V20	1.56	2.66	3.08	2.83	2.57	2.85	9.779	.000
V21	4.10	2.28	3.32	2.95	3.94	3.23	17.063	.000
V22	3	3	3	3	4	3	7.352	.000
V23	3	1	4	2	5	4	73.671	.000
V24[3]	2	2	2	2	2	2	1.424	.216

The remainder of the paper assesses these results from the application of the cluster analysis technique in the light of the previously outlined consumer behaviour matrix. As a first step in this analysis, each of the clusters was plotted on to the matrix (see Figure 6.2).

4 Research results

This section provides a discussion of the research results and relates the main characteristics of each cluster to the two dimensions of consumer involvement and consumer confidence, as shown in Figure 6.2. As mentioned earlier, deliberation and talking to somebody were taken as indicators of consumer involvement. High levels of involvement were also regarded as being indicative of consumers who had a predisposition for using either branch networks, home visits or the tele-

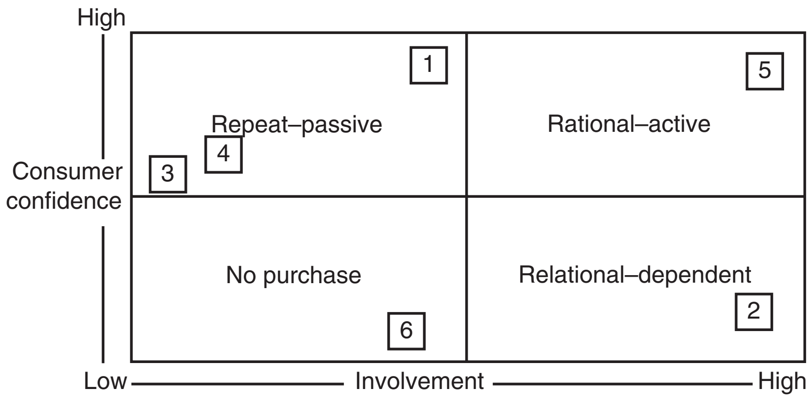


Figure 6.2 Six-cluster solutions and consumer behaviour matrix.

phone. However, in view of the emphasis on 'talking to somebody', these variables were not interpreted as being indicative of consumers who used, or would like to use, the Internet. Similarly, knowledge and understanding, and certainty were regarded as indicators of consumer confidence. High levels of confidence were also interpreted as being indicative of consumers who were inclined to use remote forms of banking such as the Internet or the telephone. Evidence of a willingness to switch financial providers was interpreted as being indicative of rational–active behaviour and unwillingness to switch was regarded as being associated with either repeat–passive or relational–dependent behaviour.

4.1 Cluster one

The involvement of consumers in cluster one was moderately high. They exhibited the lowest levels of deliberation of all the clusters for three of the four products under consideration, but felt that it was fairly important to talk to somebody. In contrast, however, confidence levels were very high. The respondents believed that they possessed a good knowledge and understanding of each of the financial products and exhibited the strongest disagreement with the assertion that they felt unsure when making purchases. In terms of switching behaviour, this group of customers had not considered switching financial products in the past two years. In fact, this group revealed the lowest levels of switching behaviour of all the clusters for each of the financial products under consideration. On the basis of these characteristics, cluster one is defined as *repeat–passive* in Figure 6.2.

High levels of confidence combined with moderate levels of involvement and a significant reluctance to switch financial providers suggests that these consumers might approximate most closely to the 'behaviourally loyal' bank customer with a preference for branch networks. As such, they may be indicative of consumers who regard the time and transaction costs associated with 'getting involved' in financial affairs, (albeit their own) or changing financial providers, as too high. Alternatively, these behaviour characteristics could be indicating that these customers are highly satisfied with their financial provider.

What may be significant, however, is that this group, which might be representative of the traditional bank customer, was the second smallest cluster in the analysis, comprising only 13.1 per cent[4] of the respondent sample. The study also sheds light upon the socio-demographics of this cluster, which possessed the highest annual household income of the six clusters, ranging from £40,000 upwards. Group members were also aged between 36 and 45 years old and possessed a professional qualification.

4.2 Cluster two

This group of consumers expressed agreement for both the importance of deliberation and, in particular, with the need to talk to somebody when selecting financial products. The group, therefore, exhibited a high degree of involvement. In contrast, however, confidence levels were low. The respondents in this cluster exhibited the lowest levels for knowledge and understanding of all the groups except for cluster six. Similarly, in terms of being unsure when selecting financial products, this group had the second highest level of uncertainty, with cluster six – once again – exhibiting the highest level. In terms of switching behaviour, although not as extreme as cluster one, this group had not considered changing financial providers. On the basis of these characteristics and by the fact that this group exhibited the highest propensity to talk to somebody when selecting financial products, it is defined as *relational–dependent* in Figure 6.2. Respondents considered it important to establish a relationship with either financial providers or third parties and this appears to have had an influence on their subsequent switching behaviour. This type of behaviour is interpreted as being indicative of consumers who prefer branch networks or home visits rather than the telephone or other remote delivery channels.

The cluster comprised 23 per cent of the respondents and, as such, was the largest segment of the sample population. In terms of socio-demographics, it possessed the lowest annual household income, ranging from £10,000 to £19,999. Group members tended to be aged between 36 and 45 years old and were educated to O or A level standard.

4.3 Cluster three

Cluster three exhibited low levels of involvement. The respondents expressed agreement on the importance of deliberation for insurance, credit- and investment-based products, but disagreed that it was important when changing a current account. They strongly disagreed, however, that it was necessary to talk to somebody when selecting financial products. In fact, this group exhibited the lowest scores for the importance of talking to somebody of all the clusters for each of the four products. In terms of confidence, however, the group was fairly confident.

They believed that they possessed a high level of knowledge and understanding of financial products and did not feel unsure when purchasing such products. In terms of switching behaviour, this group also had the highest propensity to switch.

When compared with the other groups, cluster three falls somewhere between the high confidence levels of groups one and five and the low levels of groups two and six. The group also has the lowest levels for involvement of all the clusters. This is because of the strong disagreement with the need to talk to somebody when choosing financial products. This factor, combined with the very high willingness to switch financial products, suggests that they might be using direct delivery channels, such as the telephone or the Internet, in their search and buy procedures relating to financial products. In this respect, this cluster is indicative of a group of consumers who although they appear to fit into the *repeat–passive* quadrant in Figure 6.2, are clearly distinct from consumers in cluster one. They are not as confident and they are not as involved but they express a much greater propensity to switch financial provider.

The cluster comprised 17.6 per cent of the consumers in the sample and was the third largest group. They possessed an annual household income of between £20,000 and £29,999. Significantly, they were the youngest of the six clusters, aged between 26 and 35 years old, and were educated to first-degree level. In contrast to cluster one, therefore, consumers in cluster three are both younger and on a lower annual income. This could be important in explaining the different behaviour patterns, for example younger people generally have less disposable income and less experience of finance and the purchase of investment products compared with their older counterparts. In general, their banking needs are predominately typified by current accounts, basic insurance and possibly credit-based products which demand less consumer involvement and are more amenable to switching compared with investment-based products. In this respect the behaviour of this group of consumers is probably more compatible with remote delivery channels. What is potentially important about these findings, however, is that they suggest that the younger consumer's propensity to use the telephone and possibly the Internet may be a function of their financial requirements as well as their acceptance of new technology.

4.4 Cluster four

Group four also exhibited characteristics which are readily associated with *repeat–passive* behaviour. Levels of involvement were low, but not quite as low as group three, with respondents placing slightly more emphasis on deliberation and the need to talk to somebody. The consumers in group four were also slightly more confident than those in group three, but much less inclined to switch financial products. The cluster was the smallest group, comprising 11.9 per cent of the sample population. The socio-demographic characteristics of this group revealed that it possessed the second lowest income of the six clusters, between £10,000 and £19,999. Moreover, respondents were aged between 26 and 35 years old and had the lowest educational qualifications of the six clusters, with O levels or equivalent qualifications.

It is interesting to reflect that both clusters three and four consist of relatively young consumers and that they might evolve into the sort of mature groups depicted by either cluster five (discussed below) in the rational–active quadrant or cluster two shown in the relational–dependent quadrant in Figure 6.2. For example, consumers in cluster three, like those in cluster five, exhibit a high willingness to switch financial providers, a similar level of annual household income and a comparable level of education. As consumers in cluster three mature and realise their career potential, as indicated by their educational attainments, there is a distinct possibility that they will gain in confidence and become more involved in financial affairs, thereby exhibiting a behaviour profile not too dissimilar to cluster five. Similarly, consumers in cluster four who, like consumers in cluster two, exhibited a reluctance to switch financial products, had low annual incomes and a fairly basic level of education, may eventually see their levels of confidence decline and attempt to compensate for this by developing close relationships with their financial providers. As such they might eventually migrate into the relational–dependent quadrant and become identifiable with consumers in cluster two.

4.5 Cluster five

Cluster five exhibited the highest levels for involvement of all the groups. Respondents recorded the highest scores for deliberation for each of the four products. Similarly, the importance of talking to somebody recorded the highest scores of any other group for credit and investment-based products. However talking to somebody was not as important when purchasing a current account or an insurance-based product. This suggests that the nature of the financial product, specifically its complexity, might influence the choice of delivery channels. Confidence was also high (the second highest score to cluster one). Consumers strongly agreed with the assertion that they had both knowledge and understanding of all four financial products and consequently disagreed that they felt unsure when purchasing them. In addition the group revealed a fairly high level of switching behaviour. In fact, for insurance-based products the cluster had the highest score of all the groups.

This group of respondents, which comprised 19.7 per cent of the sample population, was the second largest cluster. It combines a very high level of involvement with high levels of confidence and switching behaviour. In this respect, it corresponds to the *rational–active* mode of behaviour depicted by the top right-hand quadrant in Figure 6.2 and is conducive to the usage of telephone and even Internet banking for certain basic forms of financial product. The socio-demographic characteristics of this group also show that the respondents possessed income of between £30,000 and £39,999 and that they were the oldest of the six clusters, aged between 46 and 55 years old, with professional qualifications.

4.6 Cluster six

The consumers in cluster six had fairly low levels of involvement. They did deliberate when considering the purchase of financial products, but the levels of deliberation were relatively low, especially for investment products. With regard to speaking to somebody, the respondents believed that it was fairly important for investment- and credit-based products, but less important for current accounts and insurance products. In terms of confidence, this cluster exhibited the lowest level of knowledge of all the groups for each of the four products. Similarly, the group also revealed the highest levels of uncertainty of all the groups. This was particularly the case for insurance-, credit- and investment-based products. The group, nevertheless, shared a high willingness to switch financial providers.

In comparison with the other five groups, this cluster exhibited fairly low levels of involvement, but significantly it had the lowest recorded levels of confidence. This suggests that cluster six, which comprised 14.8 per cent of the respondent sample, should be located in the *no purchase* quadrant of Figure 6.2. A combination of fairly low levels of involvement and very low levels of confidence when purchasing financial products would appear to 'override' the group's propensity to switch and effectively stop them from making choice decisions.

The socio-demographics of cluster six revealed an annual household income of between £20,000 and £29,999, aged between 26 and 35 years old and educated to first-degree level. As such the socio-demographic factors are similar to cluster three and raises the possibility that these consumers have the potential to develop and migrate into the rational–active quadrant, either directly or via the repeat–passive route.

5 Conclusions

In this paper we have used a variety of means to validate the clusters. Arguably the real test of cluster analysis, however, is its usefulness in supporting theoretical expositions. Within the context of this paper, the derived clusters have been useful in supporting the theoretical framework of consumer behaviour in financial services.

The high non-response rate might be indicating that the majority of consumers are essentially uninterested in financial affairs and, therefore, can still be referred to as 'behaviourally loyal'. This would support the finding of Knights, Sturdy and Morgan (1994) and Watkins (1990) who found evidence to suggest that consumer inertia rather than activity and rational thinking typified the purchasing behaviour of consumers in financial services. However, a significant finding of the study is that of those consumers who responded to the questionnaire, the predominant consumer behaviour type is that associated with the relational–dependent heuristic, i.e. cluster two, which accounts for 23 per cent of the respondents. Moreover, if consumers in cluster four mature in the way suggested by the paper and eventually adopt behaviour associated with cluster two, this segment will account for approximately 35 per cent of the respondent population. At one level, this would seem to justify the relationship marketing strategies of financial service providers. However, this group of consumers is the poorest, with an average annual household income of between £10,000 and £19,900. As financial providers typically focus their relationship strategies on their more affluent customers, the size of this cluster may therefore not be reflecting the success of this strategy. In fact, the evidence suggests that an emphasis on having a relationship with a financial provider or third party is primarily driven by low levels of confidence which are a function of the relatively low levels of education associated with this group. Another interesting finding was that this group of consumers prefer branch networks rather than remote delivery channels. This suggests that the least profitable segment of the bank's customer base has a predisposition to use the most expensive delivery channel.

Another significant finding is the 19.7 per cent of the sample respondents who are represented by cluster five and associated with rational–active behaviour. If cluster three evolves in the way suggested by the paper, this group of consumers will increase to around 37 per cent of the sample and as such will emerge as the dominant respondent segment. Bearing in mind the behaviour profile of this heuristic, direct and remote delivery channels will increasingly be adopted by these consumers. This will lead to a situation where fairly affluent and profitable customers are predominately using the most cost-effective delivery channels. However, this group of consumers will pose quite formidable strategic challenges for banks and financial providers. A high propensity to switch, combined with high levels of confidence and involvement, suggest that innovations in delivery channels, value for money products, greater transparency of products, strong brand image, etc. will all be crucial in retaining and generating profitable business from this group of customers.

The smallest consumer grouping is represented by cluster one, which accounts for 13.1 per cent of the respondent sample. In some respects this group of consumers appears to conform to the description of the bank customer as being behaviourally loyal. However, as with clusters three and four, which are possibly transitory or migratory groups, cluster one does not exactly fit the behaviourally loyal customer profile. In particular these customers are very confident and although they exhibited the lowest levels of deliberation, they still, nevertheless, emphasized the importance of talking to somebody when purchasing financial products. The socio-demographic factors, especially the high income levels, suggest that they could be employed in middle-senior management positions. Their reluctance to switch might, therefore, be reflecting the demands such employment places on their time and the unacceptably high time costs associated with changing financial providers. Alternatively, these customers may already be the target of successful relationship management strategies. As they are the most affluent customer segment and, therefore, potentially the most profitable cluster with a preference for branch networks and home visits, they are certainly an obvious target for relationship management strategies. Such strategies will be important not only in realizing the full profitability potential of this group but also in helping to retain these customers in the light of increasing competitive pressure.

The final group, represented by cluster six, i.e. no purchase, represents the most difficult and strategically the most challenging for financial providers. It has the same age and income profile as cluster three and consists of consumers educated to first-degree level and yet confidence levels are extremely low and their willingness to become involved in financial decision making is not much higher. Despite this, the age and educational level of this group suggest that they have profit potential. Strategies, which emphasize trustworthiness, reliability, security, etc. might, therefore, be successful in changing the behaviour of this group and moving them into either the repeat–passive or rational–active quadrants.

In the final analysis, the low response rate to the questionnaire may be indicative of financial consumers who are still essentially behaviourally loyal. However, the research has provided tentative evidence to suggest that emerging and potentially significant consumer behaviour patterns may be represented by the rational–active and relational–dependent quadrants. Although quite different, both are probably being driven by developments in remote delivery channels, increased levels of competition and the move towards greater consumer empowerment. These forces of change appear to increase levels of confidence for some customers and make them more disposed to use remote and direct delivery channels. However, for others, an emphasis on technology and consumer empowerment might be reducing confidence levels and making them more disposed to emphasize personal relationships and branch networks. What remains to be seen is if either of these two behaviour patterns will emerge in the future as the most dominant heuristic or whether essentially inert and passive customers will continue to predominate in the financial services markets.

Notes

1 This list is not exhaustive. Depending on the type of data used additional questions may need to be addressed. For example see Punj and Stewart, 1983.

2 The value of the Kappa coefficient can range from –1 to +1 with a positive value indicating a greater than chance agreement (Cramer, 1998).

3 In light of the Kappa statistic, the inclusion of V24, despite its relative insignificance, does not seem to have caused serious deterioration of the performance of the model (Punj and Stewart, 1983).

4 Figures for size of cluster obtained from Ward's method (used henceforth for approximate size of all clusters).

References

Advertising Association (1998) *Marketing Pocket Book*. London: Advertising Association.

Bateson, J.E.G. (1989) *Managing Services Marketing – Text and Readings*. USA: Dryden Press.

Beckett, A., Hewer, P. and Howcroft, B. (2000) 'An exposition of consumer behaviour in the financial services sector', *International Journal of Bank Marketing* 18(1): 15–26.

Bettman, J.R. (1979) 'Information processing theory of consumer choice', MA thesis, University of Reading, Addison-Wesley.

Betts, E. (1994) 'Understanding the financial consumer', in P.J. McGoldrick and S. Greenland (eds), *Retailing of Financial Services*, pp. 41–84. London: McGraw-Hill.

Bloch, P. (1982) 'Involvement beyond the purchase process: conceptual issues and empirical investigation', *Advances in Consumer Research*, pp. 413–17.

Bloch, P. and Richins, M.L. (1983) 'A theoretical model for the study of product importance perceptions', *Journal of Marketing* 47: 69–81.

Brown, G.N. (1952) 'Brand loyalty: fact or fiction?', *Advertising Age* 23: 53–5.

Chase, R.B. (1978) 'Where does the customer fit in a service organisation?', *Harvard Business Review* 56(6): 137–42.

Christopher, M. (1969) 'Cluster analysis and market segmentation', *European Journal of Marketing* 3(2): 99–102.

Cramer, D. (1998) *Fundamental Statistics for Social Research: Step-by-Step Calculations and Computer Techniques using SPSS for Windows*. London: Routledge.

Doherty, N.F., Marples, C.G. and Suhaimi, A. (1999) 'The relative success of alternative approaches to strategic information systems planning: an empirical analysis', *Journal of Strategic Information Systems* 8: 263–83.

Drake, L.M. and Llewellyn, D.T. (1995) 'The pricing of bank payment services', *International Journal of Bank Marketing*, 13(5): 3–11.

Dywer, F.R., Schurr, P.H. and Oh, S. (1987) 'Developing buyer–seller relationships', *Journal of Marketing* 51 (April): 11–27.

Engel, J.F., Kollat, D.T. and Blackwell, R.D. (1968) *Consumer Behaviour*. New York: Holt, Rinehart & Winston.

Ennew, C. and McKechnie, S. (1998) 'The financial services consumer', in M. Gabbott and G. Hogg (eds), *Consumers and Services*, pp. 185–207. Chichester: John Wiley & Sons.

Fishbein, M. (1967) *Readings in Attitude Theory and Measurement*. New York: John Wiley.

Foxall, G. (1991) *Consumer Behaviour* (2nd edition). Oxford: Butterworth-Heinemann.

Hair, J.F., Anderson, R.E, Tatham, R.L. and Black, W.C. (1998) *Multivariate Data Analysis*. NJ: Prentice-Hall Inc.

Harrison, T. (1994) 'Mapping customer segments for personal financial services', *International Journal of Bank Marketing* 12(8): 17–25.

Harrison, T. (1998) 'Predicting segment membership of financial services consumers', *Journal of Segmentation in Marketing* 2(2): 45–59.

Howard, J.A. and Sheth, J.N. (1969) *The Theory of Buyer Behaviour*. New York: Wiley.

Johnson, M.P. (1973) 'Commitment: a conceptual sructure and empirical application', *The Sociological Quarterly* 14 (Summer): 395–406.

Johnson, M.P. (1982) 'Social and cognitive features of the dissolution of commitment to relationships', in S. Duck (ed.), *Dissolving Personal Relationships*. New York: Academic Press.

Katona, G. (1960) *The Powerful Consumer*. New York: McGraw-Hill.

Ketchen, D.J. and Shook, C.L. (1996) 'The application of cluster analysis in strategic management research: an analysis and critique', *Strategic Management Journal* 17: 441–458.

Knights, D., Sturdy, A. and Morgan, G. (1994) 'The consumer rules? An examination of the rhetoric and "reality" of marketing in financial services', *European Journal of Marketing* 28(3): 42–54.

Lovelock, C.H. (1983), ‘Classifying services to gain strategic marketing insights’, *Journal of Marketing* 47 (Summer): 9–20.

McDougall, G.H.G. and Levesque, T.J. (1994) ‘Benefit segmentation using service quality dimensions: an investigation in retail banking’, *International Journal of Bank Marketing* 12(2): 15–23.

McKechnie, S. (1992) ‘Consumer buying behaviour in financial services: an overview’, *International Journal of Bank Marketing* 19(5): 4–12.

MacNeil, I.R. (1978) ‘Contracts: adjustments of long-term economic relations under classical, neoclassical and relational contract law’, *Northwestern University Law Review* 72: 854–902.

Milligan, G.W. and Cooper, M.C. (1987) ‘Methodology review: clustering methods’, *Applied Psychological Measurement* 11(4): 329–54.

Morgan, R.M. and Hunt, S.D. (1994) ‘The commitment-trust theory of relationship marketing’, *Journal of Marketing* 58: 20–38.

Nicosia, F.M. (1966) *Consumer Decision Processes.* NJ: Prentice-Hall Inc.

Normann, R. and Haikola, B. (1986) *Winning Strategies in Financial Services: A Multiclient Study*. Helsinki: The Service Management Group Oy.

Norusis, M.J. (1988) *SPSS-X Advanced Statistics Guide*. Chicago: SPSS Inc.

Norusis, M.J. (1997) *SPSS 7.5 Guide to Data Analysis*. NJ: Prentice-Hall Inc.

Parsons, T. (1951) *The Social System*. New York: The Free Press.

Punj, G. and Stewart, D.W. (1983) ‘Cluster analysis in marketing research: review and suggestions for application’ *Journal of Marketing Research* 20: 134–48.

Saunders, J. (1994) ‘Cluster analysis’, *Journal of Marketing Management* 10: 13–28.

Scitovsky, T. (1976) *The Joyless Economy*. Oxford: Oxford University Press.

Shostack, L.G. (1977) ‘Breaking free from product marketing’, *Journal of Marketing* 41(2): 73–80.

Simon, H.A. (1957) *Models of Man*. New York: John Wiley.

Social Trends (1999). London: The Stationery Office.

Storbacka, K. (1994) *The Nature of Customer Relationship Profitability: Analysis of Relationships and Customer Bases in Retail Banking*. Helsingfors: Swedish School of Economics and Business Administration.

Thibaut, J.W. and Kelly, H.H. (1959) *The Psychology of Groups*. New York: John Wiley.

Tuck, M. (1976) *How Do We Choose*. London: Methuen & Co. Ltd.

Watkins, T. (1990) ‘The demand for financial services’ in C.T. Ennew, T. Watkins and M. Wright (eds), *Marketing Financial Services*. Oxford: Heinemann.

Weber, M. (1949) *The Methodology of the Social Sciences.* New York: The Free Press.

Williamson, O.E. (1975) *Markets and Hierarchies: Analysis and Anti-Trust Implications*. New York: The Free Press.

Williamson, O.E. (1985) *The Economic Institutions of Capitalism.* New York: The Free Press.

Woodruff, R.B. (1997) ‘Customer value: the next source for competitive advantage’, *Journal of the Academy of Marketing Science* 25(2) Spring: 139–144.

Acknowledgements

The author would like to thank all those consumers who took part in the research, and the financial assistance provided by NCR's Knowledge Laboratory, which made this research possible.

7 Employment perspectives in the German financial services industry and the impact of information technology*

Ralf-Henning Peters and Peter Westerheide

Abstract

In this paper we assess the determinants and perspectives of employment in the German financial services industry. We concentrate on the impact of rising use of information technology and on the composition of employment concerning different skill levels. Our approach is twofold. We first estimate employment effects by an econometric analysis. We use annual data for the time period 1978 until 1998. Second, we conduct two surveys among financial market industry experts. The results show that employment in the financial services industry will probably decrease in the coming five years. The decrease will, however, be moderate. Primarily unskilled workers are harmed by the increasing use of information technology. Labour demand for this group is expected to decline further whereas employment of highly skilled workers is expected to increase.

I Introduction

The labour markets of Germany and of many other European countries are still characterized by high unemployment rates. While in the past high unemployment was mainly a symptom of the manufacturing industries, recent discussions focus on lay-offs in the services sector as well. In this study we assess the determinants and perspectives for employment in the German financial services industry.

Employment in this industry rose until the mid-1990s and has fallen thereafter. The output, however, has steadily increased during the last two decades. Moreover, employment of different skill groups has clearly changed during the last decades. Whereas employment of unskilled workers has been decreasing, employment of university graduates has been continuously increasing. Several factors may have influenced this development. One possible factor that is widely discussed in the public domain and that we concentrate on in this study is the use of new information and communication technologies (ICT).

Innovations in computer-related technologies have particularly affected work methods in the financial services industry. Current statistics show that workers in this industry use computers more than in any other industry. In Germany in 1997 the work of almost 97 per cent of all employees in the financial services industry

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

was computer-based. In 1979 it was only 31 per cent. Investments in ICT in the financial services industry rose strongly in the last two decades.

In this study we attempt to give some insights into the impact of the increasing use of ICT on employment in the German financial services industry. Only few studies have done this for the banking and insurance industry before. Some results of these studies are summarized in Section II of this paper. A novel feature of our paper is that we distinguish between three skill levels.

According to the so-called skill-biased technical change hypothesis, the impact of new ICT is not neutral concerning various skill levels. Several studies on the manufacturing as well as on the service sector have shown that the increasing use of ICT favours the employment of highly skilled workers whereas it harms the employment of unskilled workers. As far as we know, no such studies have been conducted for the banking and insurance industry.

Our approach is twofold. We present an econometric analysis with which we try to identify the determinants of employment of various skill groups. We concentrate on the impact of ICT on labour demand. In order to assess the employment perspectives in the German financial services industry we conducted two surveys among financial services industry experts.

The results of our study show that the decline of employment in the financial services industry is expected to continue in the coming five years. Primarily unskilled workers are expected to be harmed. Demand for highly skilled workers and employment in non-bank and non-insurance institutions that also offer financial services, however, is expected to rise.

The outline of the paper is as follows. In Section II we discuss the bygone development of employment in the financial services industry and possible factors which may have influenced this development. Previous studies on this topic are sketched as well. The econometric analysis is outlined in Section III and results of the surveys are presented in Section IV. Section V gives the conclusions. Appendix 7A includes tables and figures illustrating the points dealt with.

II Determinants of the employment performance

About 1.25 million people work in the German financial services industry (according to data from 1997). In West Germany there are about 1 million employees and 0.1 million self-employed. The number of employees in the banking and insurance industry rose continuously from 1978 until 1993/4 when it began to decrease. From 1993/4 to 1998 dependent employment declined by about 3.4 per cent. Developments were similar in the banking and in the insurance industries, but they were different for the differently skilled employee groups.

Throughout this paper we distinguish between three skill groups of employees defined as follows. ‘Unskilled’ workers do not have a completed educational or vocational degree. ‘Medium skilled’ (or simply ‘skilled’) workers do have a completed educational or vocational degree but they are not graduates from a polytechnic or a university. Highly skilled workers are graduates from a university or a polytechnic (university of applied sciences).

The number of unskilled workers decreased from 1978 to 1998 by about one third whereas the number of highly skilled employees rose by about 350 per cent. The biggest skill group, with a share of about 75 per cent, was the group of medium-skilled employees. Their number rose from 1978 to 1993/4 and fell thereafter.

Some authors claim that the observed skill upgrading is due to the growth in educational levels and therefore a supply side rather than a demand side phenomenon. However, calculations based on the German Socio-Economic Panel show that the proportion of polytechnic and university graduates who are over-qualified for their position has been very stable over the past decades (Falk and Seim, 1999).

The total number of self-employed people in the banking and insurance industry rose from 1989 until 1997 from 47,000 to 66,000. In non-bank and non-insurance institutions that also offer financial services, the number of employees and self-employed people rose from 1989 to 1997 from about 75,000 up to 90,000. Thus, the self-employed people and employees in non-bank and non-insurance institutions that also offer financial services faced an employment development that was different from the employment development which dependent employees in the banking and insurance industry faced. However, since the absolute number of the former was relatively low, the positive development could not compensate the decrease in the number of the latter.

To summarize the past development of employment in the financial services industry it can be said that there is a skill upgrading and an overall reduction of dependent employment in the banking and insurance industry. Furthermore, there is an increase in the number of self-employed people and the number of people working in non-bank and non-insurance institutions that also offer financial services.

Potential determinants of employment

Innovations

Innovations in products and processes may induce increasing employment as well as originate potential for labour cost reductions. To give an example of the positive impact stemming from new products, consider the creation of financial engineered products which require highly skilled specialists. Disregarding other effects, the increasing demand for new products also leads to an increase in the demand for employees both in the distribution and in the back office. The increasing trend towards equity financing and the development of the *Neuer Markt* has added demand for highly specialized consultants in the area of corporate finance. Simultaneously a large potential for cost-saving measures exists. An example of this is the area of payment settlements where large reductions in costs have been achieved through automation. Banking products are increasingly distributed by electronic means (direct banking). Within electronic self-service, tasks are increasingly performed by the customer. Another area for potential cost cutting is telephone banking. These technologies are enhanced by developments in digital telephony and digital TV. An important aspect for future success is the implementation of digital signatures along with legal security.

Information and communication technology

Strongly correlated with innovations is the use of information and communication technology, which has increased enormously in the financial services industry. The percentage rate of German clerks in the financial services industry doing computer-aided work rose from 1979 to 1999 from about 31 to almost 97 per cent. In the US the number of automatic teller machine transactions tripled from 1987 to 1996. Unfortunately data for such a long time period is not available for Germany, but the fact that the turnover from electronic cash in Germany tripled from 1994 to 1998 indicates that a similar development took place.

In the banking industry, real investment in ICT rose from 1974 to 1998 from 200 million to more than 4,800 million DM (at prices in 1991).[1] In the insurance industry investment rose in the same time period even more – from 31 million DM to almost 1,200 million DM.[2] In the banking sector the increase was continuous until 1992. After 1992 the average growth rate decreased and investments went up and down. In the insurance sector investments increased until 1995 and remained at their high level from that time. The data for the real capital stocks, which is provided by the German Federal Statistical Office, however, does not reflect these strong increases. In both industries they increased by about 140 per cent from 1974 until 1997 and decreased slightly thereafter.

Output

Usually, a higher output is accompanied by higher employment. Depending on the returns to scale, employment may rise proportionally or under- or over-proportionally. The measurement of output in the banking and insurance industry is, however, difficult. In the literature on banking two approaches are used. One is the intermediation approach and the other the output approach. In our econometric investigation we use gross value added as a proxy for the output of the banking and insurance industry. This is the value of production minus purchases from external suppliers. As shown in Figure 7A.1 (see Appendix 7A) the real value added has risen for the last 25 years. In the banking industry there was a large increase during the period from 1993 until 1998. This stands in sharp contrast to the decrease in employment during the same period.

Wages

The wage rate and the intensity of competition are further factors which might influence labour demand. Real labour costs depend on gross nominal wages, social security contributions and producer price indices. Figures 7A.2 and 7A.3 (see Appendix 7A) show the development of the real cost of labour for various levels of qualification, based on the nominal annual income, the additional costs of labour, and the price index for producers. The different paths of the cost of labour in the commercial banking and insurance sectors are due to the different development in the price index for producers. The price index for the insurance business has increased by 38 per cent in the years from 1991 to 1998. At the same time the index for the commercial banking sector has decreased by 15 per cent. This explains the sharp increase in the banking sector at the end of the 1990s.

Competition

The trend of earnings in German and European commercial banking is marked by a sharp increase in competition. Reasons for this can be the increasing trans-border activities of banks as well as the market entry of low-priced specialized service providers such as discount brokers, direct banks and direct insurance. In both cases key forces of change are the reduction of market entry costs through the use of electronic channels of distribution and the closely related automation of processes. Beyond this the Internet has increased transparency of prices and has thereby increased the pressure on the high-price service providers.

Also, the insurance industry has come under price pressure in recent years. Similar to the developments in banking, internationalization and additional competition through direct channels of distribution are placing the established companies under pressure. Life insurance providers in banks and investment funds especially are facing increased competition from private retirement funds.

Previous empirical findings

Several strands of empirical research already provide some findings on the importance of the determinants of employment sketched above. Efficiency analyses in the banking sector represent one potentially meaningful wing of empirical research. Another relevant direction consists of empirical analysis concerning the relations between ICT and employment.

The results of efficiency analyses hint in different respects at the future employment perspectives:

- *Inefficiency*: A considerable extent of average inefficiency would indicate that future output growth would be managed not with higher employment, but with more efficient allocation of existing personnel. Growing competitive pressure on profit margins would enforce this development.
- *Economies of scale*: If increasing returns of scale can be realized on the average, future employment will not grow in the same proportion as output. Additionally, the question arises if mergers and acquisitions can yield appreciable efficiency gains.
- *Labour costs*: In the efficiency literature, cost functions have frequently been estimated. These papers offer some insights into the relationship between labour input and labour costs. Besides, substitution possibilities between capital and labour have been analysed by estimation of cross-price elasticities.

Due to the greater availability of microdata most of the existing studies refer to the US banking sector (surveys are provided by Tichy, 1990; Berger and Humphrey, 1997 and Ashton and Hardwick, 2000). The outcome of these studies is generally positive economies of scale for small banks and diseconomies of scale for large banks. However, these advantages and disadvantages of scale are comparatively small. Technical inefficiencies play a more important role. For Germany only a few analyses have been carried out (for Bavarian cooperative banks by Lang and Welzel, 1994, 1995a, 1995b, 1996; for a larger sample of savings banks, cooperative banks and commercial banks by Lang and Welzel, 1997a, 1997b, 1998; and for savings banks by Hanow, 1999 and Porembski, 2000).

These analyses for the German banking sector point to considerable average inefficiencies, which – positively interpreted – can be seen as efficiency reserves (e.g. Lang and Welzel, 1995b, 1997a, 1998). Squeezing out these reserves when competition gets fiercer would restrict future labour demand. The existence of economies of scale indicates a future under-proportionate growth of labour demand as well. However, the economies of scale are generally rather weak. External growth via mergers and acquisitions is likely to be accompanied by cost disadvantages, as long as the number of branch offices of the merged banks will not be reduced (Lang and Welzel, 1999). These results also hint at decreasing – or at least more slowlyincreasing – labour demand in the banking sector, if the current trend towards consolidation continues to prevail. Cost elasticities prove to be negative for all factors; cross-price elasticities are positive and point to substitutability between capital and labour (Lang and Welzel, 1998).

Analyses concerned with the effects of ICT on employment yielded differing results. Comprehensive empirical studies, however, seem to confirm that employment in the financial sector shows a declining trend to be explained – *ceteris paribus* – by technological progress (for Germany see Porembski, 2000 and Hanow, 1999; for the US Craig, 1997; for eighteen OECD countries Kanellopoulos, Tsatiris and Mitrakos, 1999). Analyses concerning substitutability between labour employment and ICT investment indicate that substitution is generally possible. In the short run, however, it is restricted by narrow limits (Porembski, 2000).

As a further trend, technological change is paralleled by changes in the skill composition of employees (skill-biased technological change). Several studies (e.g. Katz, 1999; Bresnahan, Brynjolfsson and Hitt, 1999; Allen, 1996) point to this relationship between ICT investment and skill structure. However, up to now no analyses have been pursued particularly for the financial sector, with the exception of Stymne, Löwstedt and Fleenor (1986).

These results lead us to suppose that future employment in the financial sector will possibly not grow. This presumption is supported by high average efficiency reserves, by economies of scale and by automation through investment in ICT.

Increasing deployment of ICT is likely to be accompanied by a skill-biased technological change which, however, has not been confirmed in the financial sector up to now. The following econometric analysis helps to fill this gap.

III Econometric estimation of a labour demand model

In order to estimate the impact of ICT on the demand for labour in the financial services industry, it is advantageous to use a long time period. The available annual data for West Germany cover the period 1978 to 1998. Since the number of employed people used to change differently among various skill groups, we differentiated by the three skill levels – unskilled, skilled and highly skilled workers.

Our econometric model of the demand for heterogeneous labour is based on the assumption of cost minimization and on a flexible specification of the cost function. For a given output level the assumption of cost minimization is equivalent to the assumption of profit maximization without any assumptions about the product market. The cost function is dual to the corresponding production function and contains all economically relevant information on the production technology prevailing in a particular firm or industry.[3] The labour demand functions for the three skill groups are derived from the cost function and are conditional factor demand functions with respect to the output. For a given output level and exogenously given factor prices, the cost-minimizing demand for labour input i, where

$i \in \{US, MS, HS\}$,

US = unskilled, MS = medium skilled, HS = highly skilled,

can be derived by differentiating the restricted cost function with respect to factor i's wage, w_i (Shephard's lemma; see, e.g., Varian, 1992: 74).

The three factor demand equations form the factor demand system. The equations are homogeneous of degree zero in all input prices and non-increasing in own input prices. The factor demand equations also imply certain symmetry restrictions, which must be fulfilled for the cost function to provide an economically meaningful description of technology.

To estimate the factor demand system we have to specify a functional form of the restricted cost function. A popular specification of the production technology is based on the so-called generalized Leontief cost function (see, e.g., Berndt 1991, Chapter 9).[4] The advantage of this cost function is that while it puts only quite general regularity conditions on the functional form of the cost function, it nevertheless implies a system of linear demand equations, which is relatively easy to estimate. However, we will also assume that the cost function is homothetic and homogeneous of order one in the level of output implying constant returns to scale in production. This assumption has the advantage of considerably simplifying the estimation of the labour demand equations by reducing the number of parameters to be estimated because all interaction terms between output and all other variables of the cost function drop out. This is meaningful for our estimation since we have only a very restricted number of observations.

We estimate the Diewert cost function for both the banking and the insurance industry in each period t (t = 1978...1998) depending on real value added, y^r, wages, w_i, real net capital stock, x^f_1 and either real investments in or the stock of ICT, x^f_2.[5] Investments and stocks of ICT are modelled as quasi-fixed variables.[6]

Alternatively, the net capital stock is replaced by a linear time trend t that should account for the technological change.

Formally, the cost function used here is given by

$$C(w, x^f, y^r, s, t) = y^r_{s,t} \times \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \beta_{ij} (w_i w_j)^{1/2}_{s,t} + \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{ij} \left(w_i x^f_j \right)_{s,t} \right.$$

$$\left. + \sum_{i=1}^{n} \delta_{i1} (w_i)_{s,t} \times t + \sum_{i=1}^{n} \delta_{i2} (w_i)_{s,t} \times t^2 \right] \quad (1)$$

where C is the real total variable cost, s is the industry (s = 1: banking, s = 2: insurance), and $\beta_{ij}, \gamma_{ij}, \delta_{ij}$ are unknown parameters.

Applying Shephard's lemma to the above cost function, we derive the conditional (with respect to output and the quasi-fixed factors) demand equations for the variable labour inputs x_i^v, $i \in \{US, MS, HS\}$: $(x_i^v)_{s,t} = (\partial C(\cdot)/\partial w_i)_{s,t}$. The division of these factor demands by real value added yields a system of input-output coefficients of the following form:

$$(\pi_i)_{s,t} \equiv \left(\frac{x_i^v}{y^r} \right)_{s,t} = \sum_{j=1}^{n} \beta_{ij} (w_j / w_i)^{1/2}_{s,t} + \sum_{j=1}^{m} \gamma_{ij} (x^f_j)_{s,t} + \delta_{i1} \times t + \delta_{i2} \times t^2 \quad (2)$$

The specification of the labour demand equations in terms of input coefficients avoids multicollinearity problems between value added and the capital stock which typically plague the estimation of labour demand equations including both as separate explanatory variables.

In principle, consistent parameter estimates can be obtained by estimating each equation separately. However, there are cross-equation restrictions implied by the homogeneity and symmetry restrictions mentioned above (see, e.g., Berndt, 1991: 461ff.). For the following reasons, these restrictions should be imposed in the estimation. First, estimated parameters in the system of equations have to fulfil these restrictions if they are to be interpreted as structural parameters of the assumed cost function. Second, the imposition of these restrictions may increase considerably the efficiency of estimation. These restrictions can be imposed by estimating the equations in (2) by a system estimator (see, e.g., Greene, 1997, Chapter 17).

Own- and cross-price elasticities for the three skill groups can be calculated from the estimated coefficients from the system of demand equations (2) (formulas are printed, e.g. in Berndt, 1991, Chapter 9). The own- and cross-price elasticities depend on relative wages, on the structural parameters of the cost function, and on the estimated input-output coefficients for the three skill groups as well as on the industry, s and on the time index, t. Therefore the elasticities vary both between and within industries.

Estimation results

At first, the system of labour demand equations (2) was estimated by the method of Seemingly Unrelated Regression with all homogeneity and symmetry restrictions imposed. Several tests show a high autocorrelation in the levels equations.[7] Therefore it seems more appropriate to estimate the equation system (2) in first differences.

$$(\Delta\pi_i)_{s,t} = \delta_{i1} + \delta_{i2}\times t + \sum_{j=1}^{n} \beta_{ij}\Delta(w_j/w_i)^{1/2}_{s,t} + \sum_{j=1}^{m} \gamma_{ij}(\Delta x^f_j)_{s,t} + u_{i,s,t} \tag{3}$$

where $u_{i,s,t}$ is an additive error term.[8] Estimation results for the system of equations in (3) with all homogeneity and symmetry restrictions imposed are reported in Table 7A.1 (see Appendix 7A).

Wage elasticities

Coefficients β_{13} and β_{23}, which refer to relative wages, are not significantly different from zero. That must not be true for the elasticities either. Standard errors have been calculated using the technique of bootstrapping. Elasticities are at most marginally significant. Own- and cross-price elasticities are shown in Table 7A.3 (see Appendix 7A). The labour demand elasticity is higher for unskilled than for skilled and highly skilled workers. This result was shown before for other industries in previous studies (see, e.g., Hamermesh, 1993). Elasticities for unskilled workers are significantly different from zero at the ten per cent level and are relatively high. Elasticities for the large group of skilled workers are –0.5 and –0.37 for the banking and insurance industries, respectively. This order of magnitude was shown in previous studies as well. However, elasticities for skilled and highly skilled workers are not statistically significantly different from zero. Thus, the hypothesis that wages do not influence the employment of skilled workers cannot be rejected at a usual level of significance. Since the elasticities for highly skilled workers are very low and not significant, it seems that wages do not influence the employment of this group of employees.

Cross-price elasticities show that unskilled and skilled workers as well as skilled and highly skilled are substitutes. Unskilled and highly skilled workers seem to be complements. This reasonable result was also shown for other industries in previous studies.

Information and communication technology

Concerning the impact of increasing usage of ICT on labour demand we get the following results. The net capital stock has a significant negative influence on the employment of unskilled workers (γ_{11}). The corresponding coefficient in the equation for skilled workers is negative but insignificant whereas the one for highly skilled workers is positive and almost significant at the ten per cent level. Thus, the capital stock is a substitute for unskilled workers and rather a complement for highly skilled workers. The impact of investments in ICT on employment is negative for all three skill groups. The *t*-value decreases with increasing skill levels. These estimations assume that the influence of ICT is the same in the banking and in the insurance industry. This assumption was tested and the hypotheses that coefficients for the two industries are the same could not have been rejected.

In the next specification we replaced the net capital stock by a linear time trend. Since we estimate first differences, the time trend corresponds to the constant. Again, in this specification the coefficients on ICT are not significantly different from zero. Only the coefficient on ICT for unskilled workers is negative. The constant is significantly negative for unskilled and skilled workers but it is positive for highly skilled workers. If the time trend is interpreted as a proxy for technological progress, the result states that progress has a negative effect on the employment of unskilled and skilled labour whereas it has no effect on the employment of highly skilled workers. These results remain true if investments in ICT are replaced by the net capital stock.

If only one out of the three variables that are used here to indicate the use of ICT, net capital stock, investment in ICT, and time trend is considered in the estimation, each is negative for unskilled and skilled and positive for highly skilled workers. The corresponding coefficients are always significant for unskilled workers but not in each case for medium and highly skilled labour.

Number of branches and summary

In one specification that was based only on observations for the banking sector, we also used the total number of branches as an explanatory variable. The coefficient is positive for all three skill groups and it is significant for the medium skilled workers. This indicates that a reduction of branches reduces the demand for skilled labour whereas it does not statistically significantly influence the demand for low and highly skilled workers.

The results show that there is a skill-biased technical change in the financial services industry. The more ICT used, the lower is the demand for unskilled and the higher is the demand for highly skilled workers. The influence of ICT on the demand for medium skilled workers is, however, less clear. On average, the influence seems to be negative, i.e. the rising use of ICT reduces the demand for medium skilled workers. But the coefficients are not always significant. Therefore, on the basis of our estimations one cannot state that there is a clear substitution effect between new technology and medium skilled labour.

IV Survey results

To assess the employment perspectives in the German financial sector and their determinants we conducted a comprehensive and detailed survey among financial markets experts in banks, insurance companies and other financial service providers. The experts were requested to give an estimate on the development of employment in their own sector for the coming five years.

In a second survey the participants of the *ZEW-Finanzmarkttest* panel were questioned on the same topics with a less detailed questionnaire. This panel consists of financial market experts employed mainly in banks, insurance companies and investment companies, and – to a smaller extent – also in finance departments of non-financial companies.

In the *Finanzmarkttest* survey 187 experts responded, the sample for the detailed survey consisted of 45 experts from banks, 12 from insurance companies, and 15 from other financial service providers.

The survey results for the employment perspectives

The results generally point to decreasing employment in banks and insurance companies (see Figure 7A.4 in Appendix 7A). For the banks, 59 per cent of the experts in the detailed survey expect a reduction of their staff in the coming five years. Forty-nine per cent of the experts, however, expect only a rather small reduction of not more than 10 per cent. Only 10 per cent predict that employment will shrink by more than 10 per cent. These findings have generally been confirmed by the results of the *Finanzmarkttest* survey, but here a higher proportion of the participants (30 per cent) are of the opinion that employment will diminish more strongly.

Concerning the insurance sector, 44 per cent of the respondents expect a reduction of the number of employees. Here also the majority anticipates a small reduction of not more than 10 per cent. The *Finanzmarkttest* participants are more pessimistic since 67 per cent of them expect diminishing employment in insurance companies.

With respect to other financial service providers the expectations differ. Almost 90 per cent of the respondents in the detailed survey expect increasing employment in the coming five years. Sixty-two per cent even anticipate a particularly strong increase by more than 20 per cent. The expectations of the *Finanzmarkttest* respondents are somewhat more moderate, but they also predict growing employment in this sector.

Skill structure

According to the experts, the skill structure will change considerably. For the highly skilled altogether growing employment is to be expected. Eighty-one per cent of the *ZEW* panel experts predict growing employment of this group in banks, insurance companies and other financial service providers. The detailed survey yielded similar results: 54 per cent of the bank experts, 60 per cent of the financial service providers and 79 per cent of the experts from insurance companies expect an increase in the employment of highly skilled employees. This will be the case in all fields of the financial business. However, the increase will mainly take place in those sectors where a high proportion of highly skilled staff is already employed. The estimates for the growth of different business fields within the financial sector confirm this assessment: in particular, those sectors which require highly skilled personnel (e.g. brokerage, mergers and acquisitions, asset management) will grow dynamically.

In sharp contrast to these optimistic predictions for the highly qualified employees stand the perspectives for the low-qualified workers. Approximately

80 per cent of the *ZEW* panel participants expect a reduction in employment of this group. Almost 60 per cent expect a decrease by more than 10 per cent.

Regarding the formal qualifications, only minor shifts are to be expected. Generally there is a slight trend towards a higher share of ICT-related qualifications.

Information and communication technologies and their impact on employment

ICT plays a central role in nearly all fields of the financial sector. Therefore it will possibly be an important determinant of employment. However, the assessment of the financial sector experts concerning this influence is heterogeneous. The members of the *ZEW Finanzmarkttest* panel were asked to give an overall assessment of the impact of ICT on employment changes. Particularly for banks and insurance companies, the answers show a great variance, while for investment companies and other financial service providers the influence is assessed as comparatively low (see Table 7A.2 in Appendix 7A).

In the detailed survey the experts were also requested to give an estimate of the impact of ICT on employment for different skill groups and business fields. For the highly skilled employees some 64 per cent of the respondents predict a positive influence on employment. The skill group with formal professional qualifications, but without university degrees, will profit from ICT use as well. On average fewer than 10 per cent of all respondents expect a reduction of employment due to ICT use in the coming five years. However, some tendencies towards ICT-induced lay-offs in the banking sector are recognizable as well (see Table 7A.4 in Appendix 7A).

V Summary and conclusion

Employment in German banking and insurance companies has been declining since the mid-1990s. Skill structures shifted in this time interval: the share of highly qualified employees has been increasing.

Results of previous studies already indicate that this trend is likely to continue. Existing efficiency reserves point to restrained growth of employment for the future. Business expansion in the financial sector is subject to moderate economies of scale. Automation tends to drive out routine work, while ICT use is likely to induce further demand for highly skilled workers.

Our study, which comprehensively analyses the volume and skill structure of labour demand for the German financial sector, confirms these presumptions: an econometric time series analysis demonstrated that different skill groups have been influenced in different ways by the increasing ICT use in the financial sector. While there is a negative impact on unskilled workers, there are complementarities between highly skilled employment and ICT investment. Furthermore, there is a significantly negative impact of rising labour costs on the employment of low-qualified employees. For the other skill groups this effect is smaller and insignificant. These results clearly point to an increasing wage differentiation as a suitable measure against lay-offs in the low-skilled employment segments.

The results of a survey among financial market experts confirm this view: in the coming five years employment in banks and insurance companies is predicted to decrease. However, this reduction most likely will not exceed 10 per cent of the current employment volume.[9] The shift in the skill structure that was observable in the past will continue to prevail. Main reasons for this continuing shift are the following: those business fields that require highly qualified labour input (asset management, investment banking, brokerage) are likely to expand. Moreover, ICT plays a dominant role in all sectors of the financial business and most respondents state a positive relationship between ICT and highly skilled employment – while there is negative relationship between ICT and the employment of less skilled workers.

Appendix 7A Tables and figures

Calculation of ICT stock and proxies

The first best indicator for the use of ICT would be its real stock in banks and insurance institutions. Apart from the problems of estimating the real stock (see, e.g., Jorgensen and Stiroh, 2000) even nominal data are not available for Germany. The lack of exact data leads us to work with several proxies for the use of ICT. As explanatory variables for ICT in our econometric investigation we apply investment in ICT, an ICT stock that is calculated from ICT investments, the general net capital stock (including ICT and non-ICT capital) and a linear time trend. In the financial services industry nowadays, about 50 per cent of all investments in equipment are investments in ICT. Since equipment is part of the net capital stock we take the latter as a proxy for the stock of ICT. The use of time trends in order to capture the impact of technological change is not unusual among researchers (for an investigation of technical change in banking see, e.g., Altunbas *et al*., 1999).

Acknowledgements

The results presented in this paper come mainly from a research project at the Centre for European Economic Research (ZEW) and the University of Erlangen-Nürnberg. Other participants were Carlo Beck, Wolfgang Gerke (project coordinator), Michael Schröder and Daniel Schäffner. The authors would like to thank the Nomos-Verlag, Baden-Baden, for the kind permission to reprint several graphs from volume 58 of the ZEW-Wirtschaftsanalysen in this article. Mr Peters acknowledges financial support by Deutsche Ausgleichsbank Bonn.

Table 7A.1 Estimation results for the labour demand system (3)

Coefficients	*Specification I* *Coeff.*	*Specification I* t-*value*	*Specification II* *Coeff.*	*Specification II* t-*value*
β_{12} (w_{US},w_{MS})	3.88	1.27	2.91	1.39
β_{13} (w_{US},w_{HS})	–1.35	–0.91	–1.24	–0.81
γ_{11} (capital)	–0.23	–3.96		
γ_{12} (ICT)	–0.26	–0.75	–0.05	–0.17
Const_{US}	–0.08	–5.44		
δ_1 (Dummy)	0.05	0.80	0.06	1.02
β_{23} (w_{MS},w_{HS})	1.32	0.43	1.93	0.76
γ_{21} (capital)	–0.28	–1.35		
γ_{22} (ICT)	–0.30	–0.23	0.45	0.71
Const_{MS}			–0.14	–2.73
δ_2 (Dummy)	0.02	0.09	0.07	0.32
γ_{31} (capital)	0.04	1.52		
γ_{32} (ICT)	–0.02	–0.11	0.00	0.00
Const_{HS}			0.01	1.33
δ_3 (Dummy)	0.01	0.27	0.01	0.26
Number of observations	120		120	
1st Equation (US)				
DW	1.56		1.78	
R^2	0.99		0.98	
2nd Equation (MS)				
DW	1.62		1.84	
R^2	0.99		0.94	
3rd Equation (HS)				
DW	1.74		1.78	
R^2	0.99		0.97	

Note: Estimation period 1980–98. *Seemingly Unrelated Regression.* R^2 refers to the levels of the variables.

Table 7A.2 Extent to which changes in employment are caused by increasing use of ICT

	Banks	*Insurance companies*	*Investment companies*	*Other financial service providers*
< 20%	19%	26%	38%	42%
20–40%	26%	24%	21%	22%
40–60%	27%	26%	23%	16%
60–80%	19%	15%	9%	12%
> 80%	9%	9%	9%	7%

Source: ZEW-*Finanzmarkttest* Panel (share of respondents).

Table 7A.3 Own-wage elasticities (mean of estimated elasticities over time)

	Banking industry		*Insurance industry*	
	Elasticity	t-*value*	*Elasticity*	t-*value*
$\varepsilon_{US,US}$	–0.96	1.71	–0.88	1.68
$\varepsilon_{MS,MS}$	–0.50	1.28	–0.37	1.27
$\varepsilon_{HS,HS}$	–0.14	0.07	–0.06	0.03

	Banking industry			*Insurance industry*		
	US	MS	HS	US	MS	HS
US	–	1.65	–0.69	–	1.51	–0.62
MS	0.35	–	0.16	0.26	–	0.12
HS	–1.47	1.61	–	–0.76	0.82	–

Note: Elasticities refer to estimation I in Table 7A.1. Calculations of standard errors via *bootstrapping*.

Table 7A.4 Negative impact of ICT use on employment in the coming five years (in %)

		Private business	*Corporate business*	*Backoffice*	*IT depts*
Without formal qualifications	Banks	38	39	45	34
	Insurance companies	0	25	20	0
	Other financial service providers	75	50	43	60
With formal qualifications	Banks	18	10	32	26
	Insurance companies	0	0	0	0
	Other financial service providers	0	16	6	10
With university degree	Banks	3	3	6	11
	Insurance companies	0	0	0	0
	Other financial service providers	0	0	0	0

Source: Detailed expert survey (share of respondents).

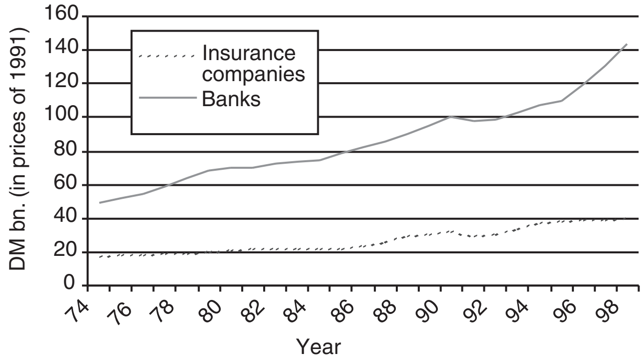


Figure 7A.1 Real gross value added.

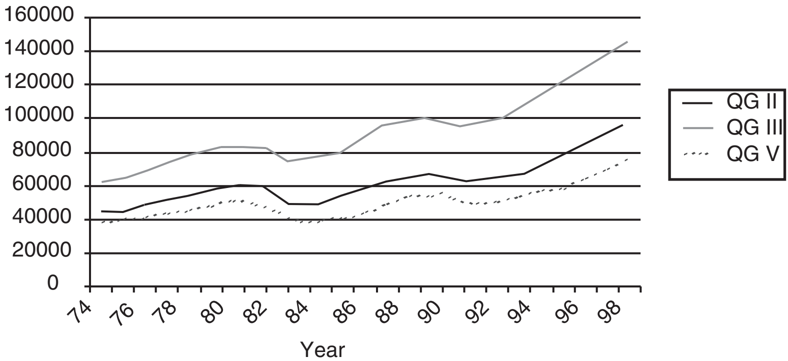


Figure 7A.2 Real labour costs in the banking industry. Own calculations based on data from the Federal Statistical Office. QG = Qualification Group, definition of the Federal Statistical Office. QG II represents highly skilled, QG III medium skilled and QG V unskilled workers.

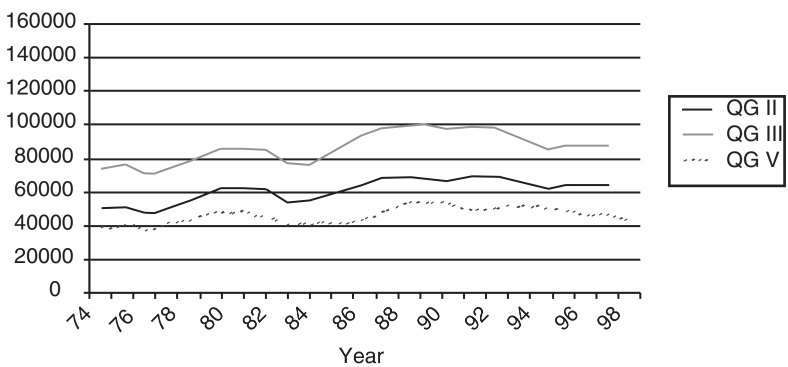


Figure 7A.3 Real labour costs in the insurance industry. Own calculations based on data from the Federal Statistical Office. QG = Qualification Group, definition of the Federal Statistical Office. QG II represents highly skilled, QG III medium skilled and QG V unskilled workers.

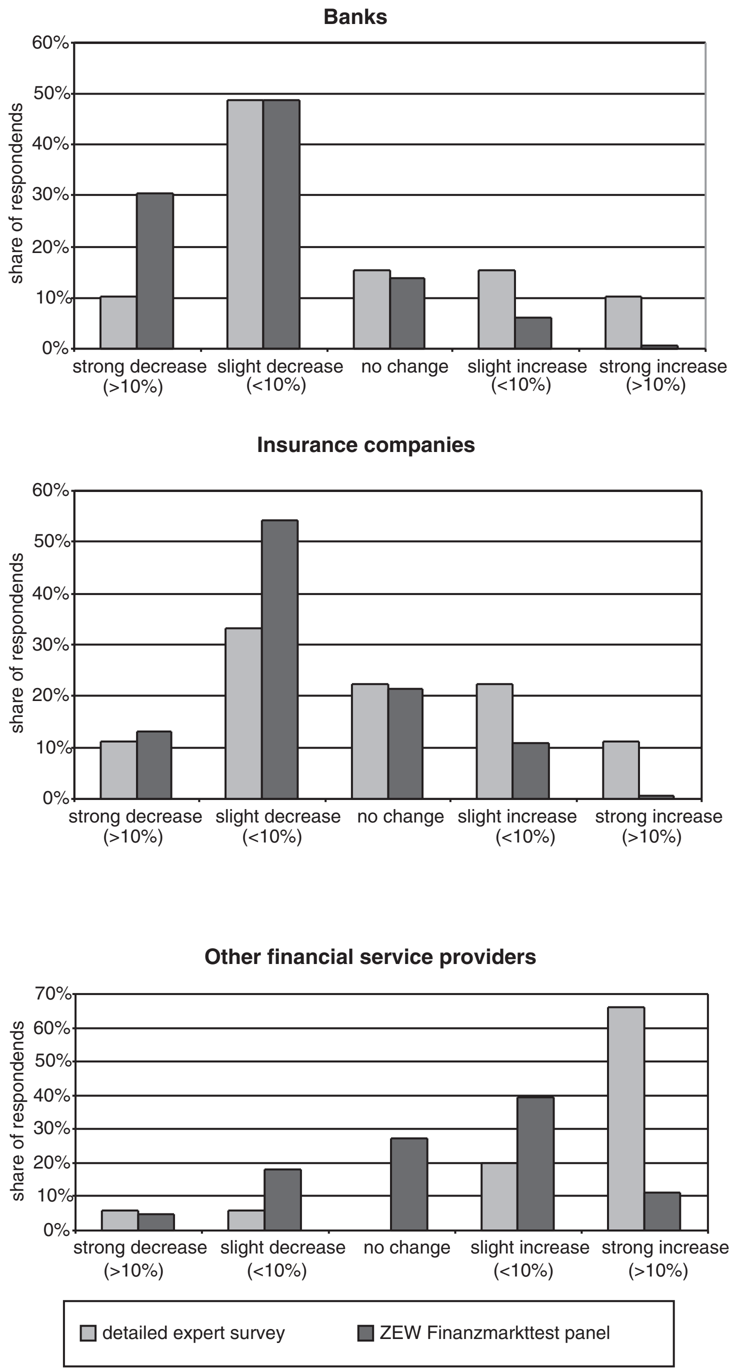


Figure 7A.4 Expected development of employment in the coming five years.

Notes

1 For details on the measurement of real ICT investments see Appendix 7A.
2 Data were kindly provided from the ifo-Institute in Munich.
3 For some standard results on the duality between production and costs (see, e.g., Varian, 1992, Chapter 4).
4 The Diewert cost function can be derived as a second-order linear approximation of an arbitrary function with continuous first- and second-order derivatives which is homogenous in prices, non-decreasing in input prices and non-decreasing in real output (see, e.g., Varian, 1992: 84f).
5 For details on the calculation of ICT variables see Appendix 7A.
6 Variable inputs can be fully adjusted within one year whereas quasi-fixed variables are not always at their long-run equilibrium due to costs of adjustment.
7 Low values of the Durbin–Watson test statistic show that the static specification with levels of the dependent and independent variables generated a high degree of serial correlation. Assuming a first-order autoregressive process, we tried to correct for autocorrelation (see, e.g., Berndt, 1991: 476ff). This procedure yielded *AR*(1)-coefficients close to 1 in each of the three equations.
8 The assumed distribution is

$$E(u_{i,s,t}) = 0, \quad E(u_{i,s,t}, \Delta(w_j/w_i)_{s,t}) = 0, \quad E(u_{i,s,t}, (\Delta x_j^f)_{s,t}) = 0, \quad \forall i,j,s,t$$

$$E(u_{i',s,t}, u_{i,s,t}) = 0, \quad \forall i \neq i', \quad E(u_{i,s,t}, u_{i,s',t}) = 0,$$

$$\forall s \neq s', \quad E(u_{i,s,t}, u_{i,s,t'}) = 0, \quad \forall t \neq t'$$

9 Calculations about the impact of the increasing demand of ICT in the German financial service industry on employment in the ICT producing industry show that employment losses in the financial industry can by far not be compensated.

References

Allen, S.G. (1996) ‘Technology and the wage structure’, NBER Working Paper 5534. Online: http://www.nber.org/papers/w5534.

Altunbas, Y., Goddard, J. and Molyneux, P. (1999) ‘Technical change in banking’, *Economics Letters* 64: 215–21.

Ashton, J. and Hardwick, P. (2000) ‘Estimating inefficiencies in banking. A survey’, *The Journal of Interdisciplinary Economics* 11: 1–33.

Berger, A.N. and Humphrey, D.B. (1997) ‘Efficiency of financial institutions: international survey and directions for future research’, *European Journal of Operational Research* 98: 175–212.

Berndt, E.R. (1991) *The practice of econometrics: classic and contemporary*. Reading, NY: Addison-Wesley.

Bresnahan, T.F., Brynjolfsson, E. and Hitt, L.M. (1999) ‘Information technology, work-place organization, and the demand for skilled labor: firm level evidence’, NBER Working Paper 7136. Online: http://www.nber.org/papers/w7136.

Craig, B. (1997) ‘The long-run demand for labour in the banking industry’, *Federal Reserve Bank of Cleveland* 33(3).

Falk, M. and Seim, K. (1999) ‘The impact of information technology on high-skilled labour in services: evidence from firm level data’, *ZEW Discussion Paper* 99–58, Mannheim.

Greene, W.H. (1997) *Econometric Analysis* (3rd edition). New York: Prentice-Hall.

Hamermesh, D.S. (1993) *Labor Demand*. Princeton University Press.

Hanow, G.A. (1999) *Die Messung des Einflusses der Informationstechnologie auf die Produktivität von Kreditinstituten mit Hilfe der Data Envelopment Analysis*. Marburg.

Jorgenson, D.W. and Stiroh, K.J. (2000) ‘Raising the speed limit: US economic growth in the information age’, OECD Working Papers, 8(87).

Kanellopoulos, K., Tsatiris, G. and Mitrakos, T. (1999) *Structural Change and Banking Employment*. Hellenic Bank Association.

Katz, L.F. (1999) ‘Technological change, computerization, and the wage structure’, Paper prepared for conference on ‘Understanding the Digital Economy: Data, Tools and Research’, U.S. Department of Commerce, Washington, DC. May 25–26. Online: http://mitpress.edu/UDE/katz.rtf.

Lang, G. and Welzel, P. (1994) ‘Skalenerträge und Verbandvorteile im Bankensektor. Empirische Bestimmung für die bayerischen Genossenschaftsbanken’, *ifo Studien* 40: 155–77.

Lang, G. and Welzel, P. (1995a) ‘Strukturschwäche oder X-Ineffizienz? Cost-Frontier-Analyse der bayerischen Genossenschaftsbanken’, *Kredit and Kapital* 3: 403–30.

Lang, G. and Welzel, P. (1995b) ‘Wachstum oder Kostenmanagement? Erfolgsstrategien für die bayerischen Genossenschaftsbanken’, *ÖBA* 8: 607–11.

Lang, G. and Welzel, P. (1996) ‘Efficiency and technical progress in banking. Empirical results for a panel of German cooperative banks’, *Journal of Banking and Finance* 20: 1003–23.

Lang, G. and Welzel, P. (1997a) ‘Größe and Kosteneffizienz im deutschen Bankensektor’, *ZfB* 67(3): 269–83.

Lang, G. and Welzel P. (1997b) ‘Non-parametric efficiency analysis in banking. A study of German universal banks’, Institut für Volkswirtschaftslehre der Universität Augsburg, Beitrag nr. 163.

Lang, G. and Welzel, P. (1998) ‘Technology and cost efficiency in universal banking. A “thick frontier” – analysis of the German banking industry’, *Journal of Productivity Analysis* 10: 63–84.

Lang, G. and Welzel, P. (1999) ‘Mergers among German cooperative banks’, Working Paper Series B nr. 99–03, Department of Economics, Friedrich-Schiller-University Jena. Online http://www.wiwi.uni-jena.de/Papers/wp-b9903.pdf

Porembski, M. (2000) *Produktivität der Banken. Untersuchungen mit der Data Envelopment Analysis.* Wiesbaden: Deutscher Universitäts-Verlag.

Stymne, B., Löwstedt, J. and Fleenor, C.P. (1986) ‘A model for relating technology, organization and employment level: A study of the impact of computerization in the Swedish insurance industry’, *New Technology, Work and Employment* 1: 113–26.

Tichy, G. (1990) ‘Bankengröße und Effizienz’, *Kredit and Kapital* 23(3): 358–88.

Varian, H. (1992) *Microeconomic Analysis* (3rd edition). New York: Norton.

8 Technology and the (re)location of financial activity

A European perspective*

*Iman van Lelyveld and Marieke Donker***

1 Introduction

Information has always played a very important role in financial intermediation. An important function of financial intermediaries is, for instance, overcoming asymmetric information problems. Another is the transformation of funds in space and time. Both these functions require banks to process information efficiently. It is therefore not surprising that the information technology revolution has brought some profound changes to the financial sector. Marketplaces have become disembodied, as for example the electronic stock market NASDAQ. More recently, the emergence of electronic communication networks (ECNs) has created virtual marketplaces. More areas of banking have been touched by IT. Backoffice operations, for example, can now be located at some distance in lower cost areas.

(Internet-based) technology has made it possible to transmit large volumes of information and to access it from remote locations at low cost. This has two possible effects on banking. On the one hand, the external market environment could change: actual markets could become virtual marketplaces. On the other hand, the production process of financial services could become even more digitized.

As this might reduce the importance of distance and place, the question arises as to what the effect could be on the location of financial activity. Will geography become irrelevant, as argued by O'Brien (1992)? Or does place still matter, but do different places matter for different financial activities, as argued by Tschoegl (2000)? In the absence of other determining factors, we would in the first case see a complete dispersion of financial activity whereas in the second case, activities would be separated geographically, leading to concentrations in a few places.

However, long-term data on Internet applications, necessary to analyse a sticky process like financial sector relocation, are not yet available. Therefore, we turn to the 'natural experiment' of the last two decades. 'Normal' communication technology has not only improved rapidly, but has become cheaper and more widely

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

** De Nederlandsche Bank, Directorate Supervision. The views expressed in this paper are those of the authors and not necessarily those of De Nederlandsche Bank. We thank Martin Hallet for his willingness to share his data set, our colleagues at DNB and the participants at the SUERF Colloquium for helpful comments on this and previous versions.

available as well. The telex, with its rather rigid infrastructure, has for instance been superseded by the fax. Voice communication has increased. Growth has not been limited to fixed lines but also mobile communication has shown impressive growth rates.[1] For financial institutions, this development offers the possibility of supplying services along different, lower cost channels. Figure 8.1, for instance, shows that the costs of a payment transaction using modern technology is significantly lower.

Although the Internet differs fundamentally in its openness and fundamental larger reach, from other channels, it is similar in the sense that it makes communication easier and cheaper. We argue therefore that developments in the past two decades will give us a first-order approximation of the effects of improved communication technology.

In this paper, we investigate the 'Geography Doesn't Matter' or GDM hypothesis, which says that the production of financial services could take place anywhere and geography has become irrelevant. To this end we analyse the location of financial activity, looking for a relocation of financial activities leading to a more even distribution of activity across space. Technology is, of course, not the only factor influencing the location of financial activities; numerous other (centripetal and centrifugal) forces play a role as well. A subject not analysed in this paper is how technology has influenced each of these other factors separately. It might be possible that some factors have become more important, whereas the role of others has diminished or has not changed at all, so that the overall effect of the various forces is unknown. Given the importance of technology and marked change in the use of technology we expect to see some change. Since we cannot adequately measure other relevant factors we would then be at a loss for a more exhaustive explanation of changes. More precisely, we focus on the GDM hypothesis.

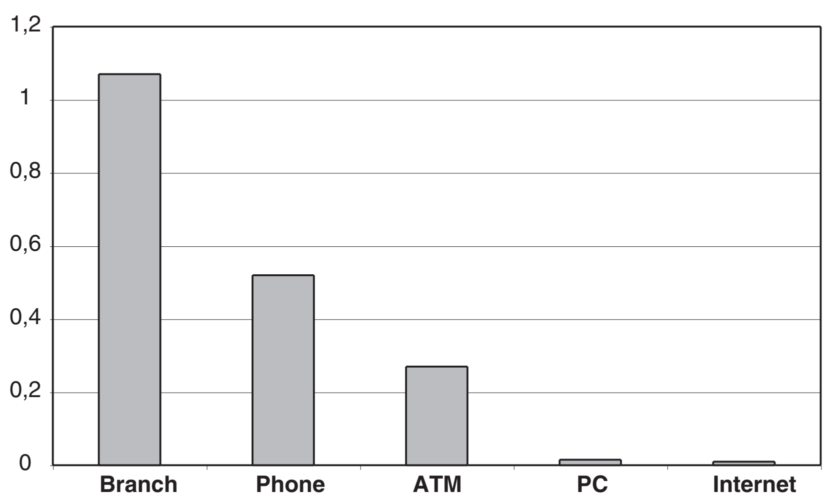


Figure 8.1 Cost of a payment transaction (in dollars).
Source: Booz-Allen and Hamilton, 1998.

Generally, financial activity is measured by market turnover, which is an appropriate measure for activity on, for instance, exchanges. However, by looking at turnover as a measure for financial activity the scope is limited to those activities that take place on markets. We take a broader perspective and analyse gross value added and employment in the financial sector. This enlarges the scope as it also includes services like, for example, financial advice. From a political economy standpoint the development of employment, an important determinant of voting behaviour, is of interest. Since relocation is a slow process we use the longest possible time series (1980–95) and we will limit ourselves to a European perspective.

In analysing the impact of technology on the geography of finance, we touch on the question of what a financial centre is. Many definitions have been provided, but there is no universally accepted definition of an international financial centre (see Dufey and Giddy, 1978; Nadler *et al.*, 1955; Kindleberger, 1974 and Scholtens, 1992). Generally, financial centres can be considered as an aggregated collection of activities which differ in many respects.[2] In this paper, we define a financial centre as a geographical concentration of activities of credit and insurance institutions. Moreover, since our data is on the regional level, the activities in what we term a ‘centre’ will not be as concentrated as is usually envisioned.

The set-up of this paper is as follows. In order to analyse the impact of technology on the location of financial activity, we first take a look at the forces that drive concentration and dispersion of these activities in theory. Then we give a description of the data, followed by an empirical analysis. In this analysis, we first examine the changes in real gross value added and then turn to the development of employment across regions. Additionally, we will, using the data on production and employment, consecutively, analyse whether production has converged to regions with an efficient workforce. Finally, some concluding remarks are given.

2 Review of the literature

The study of the spatial concentration of economic activity goes back to Alfred Marshall (1920) who presented the seminal economic analysis of the phenomenon. He identified three main agglomerating forces, which are generated by positive externalities at the level of the firm. These forces include labour market externalities, the availability of intermediate inputs, and technological as well as informational spillovers. These three forces can be applied to the financial sector as well. In the following paragraphs, we discuss the factors driving spatial concentration – and the countervailing forces of dispersion – of financial activity.

In the literature, the forces influencing the location of financial activity are divided into two groups: centralizing or centripetal forces and decentralizing or centrifugal forces (Gehrig, 1998; Walter, 1998). Whereas centripetal forces contribute to spatial concentration of activities in a few places, centrifugal forces cause geographical dispersion. The location of activities in particular places is likely to be the outcome of the interplay of these factors and is likely to be changing over time.[3]

Important *centripetal forces* are economies of scale, which, for example, can be realized in the payment system and the clearing and settlement of transactions. Kindleberger (1974) states that centralization of all payments in a single centre creates the most efficient settlement system for a given number (n) of financial centres. Without centralization, $n^x(n - 1)/2$ communication channels are needed to effect payments, whereas with a single centre $n - 1$ channels suffice. Although economies of scale increase the pressure for clustering of certain activities, these activities do not necessarily have to take place within a defined geographical area. For example, payment systems are run by computers which are part of electronic communication networks. However, the location of these computers is not relevant for achieving the desired scale economies. The same holds for currency trading, which is performed within computer networks as well (Gehrig, 1998). Access to a trading network enables a financial agent to trade from any place in the world. These examples indicate that for certain activities, concentration of participants seems to be important because of economies of scale in the transaction technology. However, the concentration takes place within one single electronic network and the particular location of activity and participants is not relevant, except for possibly regulatory reasons (Gehrig, 1998).

In contrast to these economies of scale within a single electronic network, a different situation arises when external economies are present. In that case, participants benefit from the presence of other participants at a certain location, strengthening the importance of physical proximity. We will turn to such external economies now. First, we discuss the externalities arising from the general forces identified by Marshall for the financial sector. Then, we analyse some factors which are important for financial activity in particular.

Regarding Marshall's first factor, labour market externalities, a concentration of activities in the same place allows a pooled labour market which offers a constant, deeper market for skill. This externality stimulates concentration from the demand side of labour as well as from the supply side, especially when skills are highly specialized. Employers will resort to locations where they are likely to find a large choice of workers with the special skills they require. Employees go to locations where many employers, needing such specialized skills, are established because there they will find a 'deep market' for their services. The importance of the labour market for the location choice of financial activities is underlined by the fact that finance generally requires highly skilled and often specialized labour forces. As pointed out by Kim (1990), a large labour market provides better expected job matches, which means that a large pool of labour offers more protection against idiosyncratic employment shocks. Therefore, financial agents might tend to locate in centres with large and liquid labour markets, as this may mean a quicker filling of vacancies for highly skilled labour. Rauch (1993) considers the average level of human capital as a local public good. He states that by centralization of finance in cities with higher levels of human capital, human capital externalities can be gained which may result in lower costs for the industry as well as in possibly higher quality of the services. Begg (1991, p. 338) argues that 'the availability of a labour pool is an inducement to firms seeking to relocate'.

The second point stressed by Marshall, the availability of intermediate inputs, can also be a factor contributing to the concentration of financial activity. As financial intermediaries are users of services like accounting and legal advice, locating near suppliers of such services could be advantageous since they are often needed timely. If, for example, a corporate finance deal needs to be put together quickly, it is a practical matter to have the necessary lawyers and accountants available in one place. Stuart (1975) has shown that for cases where face-to-face contact in the negotiation and purchasing process is required, a concentration of suppliers diminishes the clients' search costs and increases the market size for every supplier. The work of Gaspar and Glaeser (1996) suggests that telecommunication technologies may be a complement, or at least not a substitute, for face-to-face interactions. Recently, Porter (1998) has argued that the enduring competitive advantages in the modern, global economy arise from concentrations of, among others, related businesses, institutions and competitors' institutions.

The last force identified by Marshall, technological spillovers, states that the local spillover of knowledge among firms may give rise to concentration.[4] This may be a major factor for an industry like the financial sector, where technological innovation plays an important role. Product innovations could generate positive externalities since the diffusion of new products can take place rapidly when spatial concentration is high. This is especially the case because financial innovation is generally poorly protected by patent law.

Another particularly relevant externality arising from the concentration of financial activities is the spillover of information. Financial agents may benefit from the presence of other agents at a given site as it enables them to benefit from their information. By locating at that site, access to the information flow is maximized. The value of information, however, depends not so much on its quantity, but especially on its quality and timeliness. According to Porteus (1996), this makes face-to-face contact and other localized information sources important as means of rapid information diffusion. He recognizes that improvements in telecommunication technologies have made access to information more equal regardless of location, but argues that a distinction according to the type of information has to be made. *Standardized* financial information (like price quotations) can be transmitted quickly and at low cost over computer networks. In contrast, for the electronic transmission of *unstandardized* information, the quality may decline when the distance between the user and the source of information increases. He illustrates this with the example that a firm-specific rumour may spread rapidly through a network, but that for agents further from the source it is more difficult to confirm and interpret the information in order to exploit it usefully. 'Information has value to users only in its context, against a background in which it can be correctly interpreted' (Porteus, 1996, p. 8; see also Thrift, 1988). However, because this informational background cannot be easily standardized and transmitted across computer networks, it is more difficult to evaluate the information. Agents located near the source may have a better and quicker insight into the value of the new information, which gives them a temporary informational advantage. The importance of physical proximity to generate such a temporary informational advantage is also stressed by Gehrig (1998), who states that this is particularly important for informationally complex and sensitive products. A related argument is that because information has become more widely available, the incentive to find unexploited information has increased. This has the countervailing effect on agents to gather closer to sources of information. According to Scholtens (1992), the concentration of financial activity can be considered as a means to cheaply distribute available information as well as a means to lower the cost of acquisition of information.

Having discussed Marshall's three more general factors influencing location decisions, we now turn to some factors that particularly apply to the financial sector. Gehrig (1998) points out that for the financial sector in particular, a factor driving concentration is market liquidity. In liquid markets, individual transactions cause no significant price fluctuations whereas in illiquid markets even small transactions may cause considerable price movements. As the risk of price fluctuations is lower in liquid markets, risk-averse investors will prefer to trade in markets where liquidity is high. As a consequence, liquid markets will attract more trading volume, which gives rise to big, concentrated markets. However, thanks to technological developments like electronic trading and the recent emergence of electronic communication networks (ECNs) in the USA, physical proximity seems no longer necessary to achieve high levels of liquidity. Therefore, the argument of market liquidity is more likely to be an economy of scale within the electronic system as opposed to an externality at the level of the firm.

Also, structural changes in the financial sector have influenced the concentration of activities. The growing role of fund managers in financial markets has attracted significant amounts of capital into the hands of some big players. Fund managers prefer to be close to markets and to each other for the above-mentioned externalities (Porteus, 1996).

In contrast to the above-discussed centripetal forces, some decentralizing or *centrifugal factors* can be identified. Obviously, locating in financial centres involves high cost and congestion problems. Because of the first reason, back-office functions have increasingly been decentralized to low-cost locations. According to Rosen and Murray (1997), the decline of the share of employment in the financial sector in New York was partly caused by the relocation of back-office functions. McKillop and Hutchinson (1991) found some evidence that in the UK also a dispersion of head offices can be observed, besides the decentralization of backoffice activities.

The cost of market access constitutes another centrifugal force. This subject is analysed by Pagano (1989). As market access involves costs, investors have to decide which market to enter. In order to choose the optimal market, they will compare the expected utility of market participation net of the cost of entry. The expected utility depends, for reasons of the liquidity externality, on the participation of other agents. When access costs are symmetric, concentration will always take place in one market. However, when access costs differ across markets, fragmentation may exist. Furthermore, the spillover of information (a centripetal force) can also be used to argue that agents concentrated in one location are less able to take advantage of non-local information. When the value of non-local information increases, this could give rise to geographical dispersion. Different time zones are another factor favouring decentralization (Scholtens, 1992). Differences in time zones give rise to the emergence of financial centres, spread around the globe, among which trading is possible twenty-four hours a day (Thrift and Leyshon, 1988). Finally, a factor that influences the location choice is that financial service suppliers tend to follow their major (international) clients in order to serve them optimally (Towey, 1974; Campayne, 1992). The continuation of servicing already established client relationships has been put forward as a motive for the internationalization of financial intermediaries (Bryant, 1987). In this sense, financial activity follows real activity. Buch (2000) finds a strong and positive correlation between the foreign activities of German banks and the foreign activities of German firms, concluding that this finding supports the hypothesis that German banks follow their customers abroad. Kindleberger (1983), however, argues that there is no easy way to determine whether banks follow or lead international business. According to his opinion, banks lead and industry follows when banks are aggressive in building world networks and when industry concentrates on single projects. Under the opposite conditions banks follow and industry leads.

Besides the above-described centripetal and centrifugal forces, some more *general factors* also affect the location of financial activity. Examples include path dependency, politics and regulation. A path dependence process can be described as a non-ergodic sequence, or one in which the initial conditions determine subsequent outcomes (Porteus, 1996). In this situation, historical, locational events may have cumulative consequences in the long run. Martin (1999) refers to Krugman (1991) and Arthur (1994), who argue that because of the existence of some uncertainty in industrial location and agglomeration, several alternative equilibria are possible. Which particular equilibrium pattern of activity occurs is, to a large extent, determined by history and does not always have to be strictly superior to the alternatives. However, once the initial pattern is established it becomes rather rigid as forward and backward linkages and self-fulfilling expectations reinforce this pattern (see also Fujita and Thisse, 1996).

Politics can both facilitate as well as impede the development of a financial centre. The emergence of London as centre for the Eurodollar market serves as an example for the case in which politics clearly influenced the emergence of a financial centre. In contrast, the lack of financial centres in the Eastern Mediterranean and the Balkan, is also largely caused by politics (Jones, 1992).

The financial sector is generally considered as one of the most heavily regulated. Therefore, the regulatory and fiscal environment (tax policy and prudential supervision) affects the locational behaviour of financial agents. Scholtens (1992) distinguishes two contradictory forces which are relevant in this context. On the one hand, financial agents may take advantage of the credibility and reputation of their regulatory authorities. Agents who comply well with the strict and adequate rules of a respected regulator, may gain a reputation compared with agents which are subject to less strict rules. On the other hand, operating under lax rules and low taxes is less costly and therefore more profitable. Agents thus face a trade-off between the reputational advantage of a strictly regulated country (with higher costs) and the advantages of a less strictly regulated environment with a lower (or negative) reputation (with low cost). The choice for a location will differ per activity, depending on the relative importance of the reputation and cost dimensions.

The analysis so far has taken the macroeconomic view based on markets or the products traded in them. A different strand of the literature approaches the conglomeration question from the point of view of the individual firms. In this line of research, expanding the seminal work of Coase, *The Nature of the Firm* (1937), the question of interest is whether a financial firm should internalize some action that was previously performed in the market. In line with this sort of research are the so-called ‘Least-cost location theories’ (Weber, 1929; Isard, 1960 and Smith, 1981) in which location is primarily driven by cost factors. The largest cost element in banking is labour. Other important cost drivers are communication costs and cost of regulation. Reed (1981), for instance, shows that New York, London, Tokyo and Hong Kong have all benefited from the nearby hubbing of communication networks. Regulatory costs are difficult to estimate but are thought to be substantial and to differ widely across centres.[5] A criticism of the least-cost location theory is that it does not take into account the demand side. Optimal location is analysed separately from potential sales in the different locations. Related to this argument is the follow-the-customer motive that states that banks tend to follow the locational pattern of their major clients.

Summing up, we have described the major forces that influence the location of financial activity. Whereas centripetal forces, such as positive external economies of scale, give rise to a concentration of activities, centrifugal forces, such as high costs, cause a geographical dispersion. We now turn to our empirical investigation of the location of financial activity in Europe, in which we analyse whether it has changed under the influence of technological developments.

3 Description of the data

We analyse two measures of financial sector activity: gross value added (GVA) and employment. Since, in the data at hand, GVA is a nominal value, it is transformed into real terms using the consumption price index of the IMF International Financial Statistics (line 64). The second measure of the location of financial sector activity is how much employment is generated in a region. For each of the measures, the financial sector covers both the credit and the insurance industry.

Data, covering 1980 through 1995, are taken from the Eurostat Regio database, which provides a standardized classification of regions in Europe at various levels (NUTS). We use data on NUTS II-level – a classification with 174 sub-regions – as far as possible. If such regional data are not available, more aggregated data, on either NUTS I-level or national level, are used. However, some gaps in the data remained, which have been filled by data from national accounts or, in some cases, extrapolation. A detailed description of the data per country is given in Appendix 8A. It shows some key characteristics per country, i.e. the number of regions, and the average and the median real GVA and employment in the period 1980–95. Differences between countries seem to be substantial, ranging from a low mean of 504 in Portugal to a high 10,628 in Austria. However, part of this difference is caused by the unavoidable difference in the size of regions. In this respect, especially Austria and Sweden seem to be 'too large'. We will return to this issue in the analysis below. Data on employment cover a shorter period, i.e. 1980–92, and fewer regions. Again, these data have been taken from the Eurostat Regio database. For some regions however, data were not available. In those cases we used data from the International Labour Organization (ILO), which provides data on a national level (See Appendix 8A, Table 8A.2). The number of regions, average employment and the median are shown in Table 8.1. Countries for which only two observations were available have been excluded from the analysis. A remarkable feature is the high mean level of employment in Denmark compared with other countries.

A point to note is that if banks outsource more of their activities outside the category of 'financial services', measured activity declines although 'true' activity is unaffected.[6] However, as long as banks outsource inside the 'financial services' category no harm is done. Moreover, if outsourcing happens at an equal rate across regions the shape of the overall distribution is unchanged.

Table 8.1 Key characteristics of real gross value added and employment in the financial sector, per country, 1980–95

	Real gross value added			*Employment*		
Country	*Regions*	*Mean*	*Median*	*Regions*	*Mean*	*Median*
Austria	1	10,628	10,818	—	—	—
Belgium	11	852	483	11	12	8
Denmark	1	2,615	2,537	1	97	100
Finland	1	2,951	2,794	—	—	—
France	22	2,099	1,032	22	27	16
Germany	10	5,998	4,562	—	—	—
Greece	1	1,886	1,468	—	—	—
Ireland	1	2,465	2,174	1	37	39
Italy	20	2,196	1,129	20	20	11
Luxembourg	1	1,771	1,832	1	14	13
Netherlands	12	971	536	—	—	—
Portugal	7	504	148	5	13	7
Spain	18	1,647	780	18	17	8
Sweden	1	9,120	9,140	1	89	88
United Kingdom	11	3,748	2,056	—	—	—

Source: Eurostat Regio database.

4 Empirical analysis

The central question in this paper is whether 'geography matters'. Specifically, if technology has reduced the importance of geography, over time, the distribution of financial activity should become more even across regions. In the extreme case, distance no longer matters and the distribution of financial activity would, in absence of other influences, become completely uniform.

To analyse the impact of technology on the geography of finance, we look at two measures of financial sector location: GVA and employment. The set-up of the analysis for each variable is as follows. First, we examine the development of the distribution, estimated non-parametrically, across time. The graphical evidence is then complemented with more formal and parametric tests. Following this general examination we turn to three particular effects, discussed in the theoretical literature, that might influence location:

- *The effect of a home market:* Might the choice of location not still be driven by the size of the local market?
- *The role of financial centres:* Are financial centres governed by different 'laws of motion' than adjoining or peripheral regions?
- *The importance of regulation:* Notwithstanding any effect of changes in information technology, (national) regulation might still be such a strong influence that its effects are swamped. We thus look at the effect of the 1992 'One Market' policy package and in addition, we examine the effect of borders. In the analysis of all three aspects, we first estimate the distribution non-parametrically and then turn to more formal measures.

5 Gross value added

A crude but often used graphic to reveal distributions is the histogram.[7] A more sophisticated approach is to use a kernel density estimator.[8] The difference with a histogram is, first, that the bins are allowed to overlap and, second, that the observations in the bin are given varying weights depending on the distance from the centre of the bin. These weights are determined by the kernel function. The advantage of kernel density estimates is that they are smooth and independent of the origin. Density estimates are valuable because they are very well suited to reveal skewness, developments in the tail of the distribution, and – possibly – multimodality. At this point we want to preclude a parametric approach to capturing densities in order to allow the greatest possible flexibility. Just like in the case of the simpler histograms, the estimates produced are sensitive to the chosen bin width and, less importantly, the kernel. However, the results shown below do not change qualitatively under various combinations of kernels and bin widths.

We estimate the distribution of the data in each year in the period 1980–95. Graphical representations of the consecutive years are drawn in a single graph in Figure 8.2. Figure 8.2a is the result for the full sample while Figure 8.2b and c show the distribution of the below- and above-median subsample.

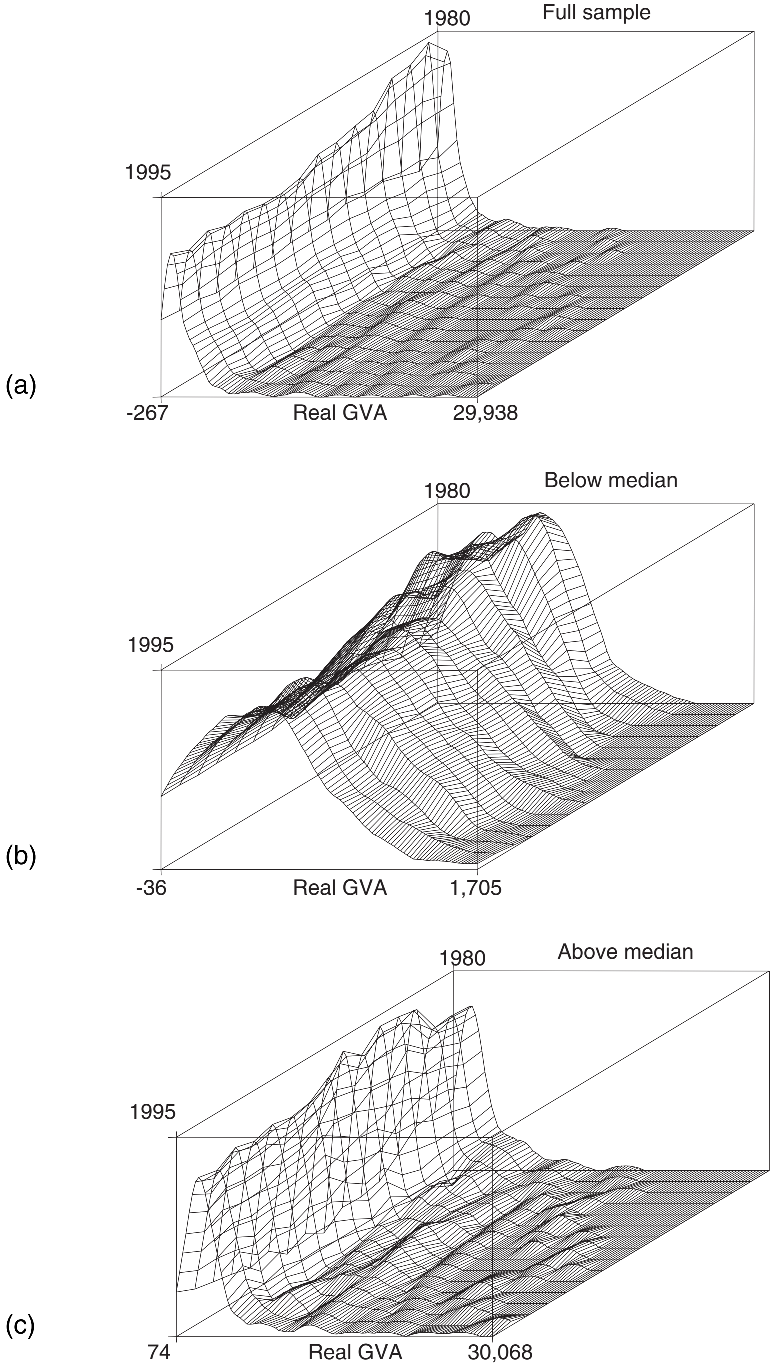


Figure 8.2a–c Kernel density estimation for GVA in European regions, 1980–95.

The estimated distribution is highly skewed, as expected, and it does not seem to be changing much over time, except for an increase in range from around 40 to 16,000, rising to 27,000.[9] This implies that the dispersion has increased a little bit. The two smaller graphs show estimates for the below- and above-median subgroups. Both distributions largely do not change, except that the distribution below the median becomes somewhat smoother.

Although the density estimates shown are very useful for discovering general features of the distributions, no stronger conclusions can be drawn. Many, more formal, measures have been suggested to capture inequality or diversity (see Cowell, 1995, for an excellent introduction). The question how 'unbalanced' a distribution is, is quite relevant in, for instance, the analysis of welfare, inequality and poverty. As we are dealing with a continuous distribution, we compute two well-known measures: the Gini coefficient and the coefficient of variation. The Gini coefficient is defined as the average difference between all possible pairs of GVA in the population, expressed as a proportion of total GVA (Cowell, 1995) and has a minimum value of 0 and a maximum of 1, where higher values mean higher concentration.[10] In the context of income distributions, the typical application of the Gini coefficient, the Gini computation, compares the actual distribution with a distribution where every individual has the same income. In our study this translates into the GDM hypothesis: as distance becomes less important the computed Gini coefficient should fall. However, as is obvious in Table 8.2, the coefficients and thus the inequality seem to be rising, not falling. This result is in line with Hallet (2000), who finds that financial services are, among sixteen other sectors, the most concentrated.

As a second measure for inequality, we compute the coefficient of variation, which gives an indication for the dispersion within the population.[11] The range of the coefficient is given by $(0, \sqrt{n-1})$ with higher values indicating more dispersion. In our sample, the computed coefficients seem to be rising.

We calculated other measures of inequality as well (not shown) but these all point in the same direction.[12] The more formal measures thus all seem to indicate that regions have become more diverse, not more uniform.

5.1 The home market

A criticism one could make of the data used is that the regions we compare are not equal. Some are larger, in terms of geographic area or population, than others. The analysis in this section shows that when we correct for a plausible 'size' measure, the results remain unchanged.

Table 8.2 Formal measures of inequality, 1980–90

	1980	*1985*	*1990*	*1995*
Gini	0.60	0.61	0.62	0.64
Coefficient of variation	1.40	1.51	1.54	1.70

Producers of financial services not only produce for the world – or even the European – market. A sizeable part of the production is for local distribution. Ideally, we would like to divide a region's production into a local and an inter-regional part in order to investigate the impact of the home market. Unfortunately, such level of detail is not available in our data. Assuming that the number of inhabitants is a good indicator of the local market, we weigh the sample by population and re-estimate the densities through time.[13] The results are shown in Figure 8.3.

When we weigh the observations with a relative population weight, the overall density becomes more concentrated. This implies that home market effect is in existence: populous regions also have large production. Both distributions, real gross value added and population, are thus similar and work in the same direction. Splitting the sample in below- and above-median does not reveal substantially different patterns compared to the sample without weights.

We also computed the population-weighted densities for the centre, adjoining and peripheral subset (shown in Figure 8.3d to f).[14] Except for the somewhat higher concentration on the left-side of the spectrum, the graphs do not differ markedly from the figures to be discussed in the next section.

5.2 Centre and periphery

A sizeable literature has emerged on the importance of hierarchy in the development of a financial sector. If rank is an important attribute of a financial centre then there might be some movement in a subsample that is concealed by larger developments in other parts of the distribution. We therefore continue the analysis on the level of centres, adjoining regions and periphery. The estimated distributions are shown in Figure 8.4.

The distribution of GVA in the seven largest centres, i.e. the regions with the highest GVA values, seems to become less peaked. This would indicate that, in the league of centre regions, similarity is decreasing. A similar movement is discernible in the distribution of GVA in adjoining and peripheral regions. Since the centre, adjoining and peripheral classification maps fall more or less directly into a classification based on population or market size, there does not seem to be a large difference due to the importance of being a centre region. Moreover, there seems to be no move of activities from centres to low-cost, peripheral areas as predicted by the 'Geography Doesn't Matter' hypothesis.

5.3 Regulation

All in all, the estimated distributions so far do not show a strong tendency towards a more even distribution of activities across space. It seems that technology has not significantly influenced the location of financial activity. However, related effects could have an impact on location choice as well. These effects could, for instance, be any significant changes in the market environment or national effects, in particular regulation.[15] We look consecutively into the effect of borders and the 1992 'One Market' policy.

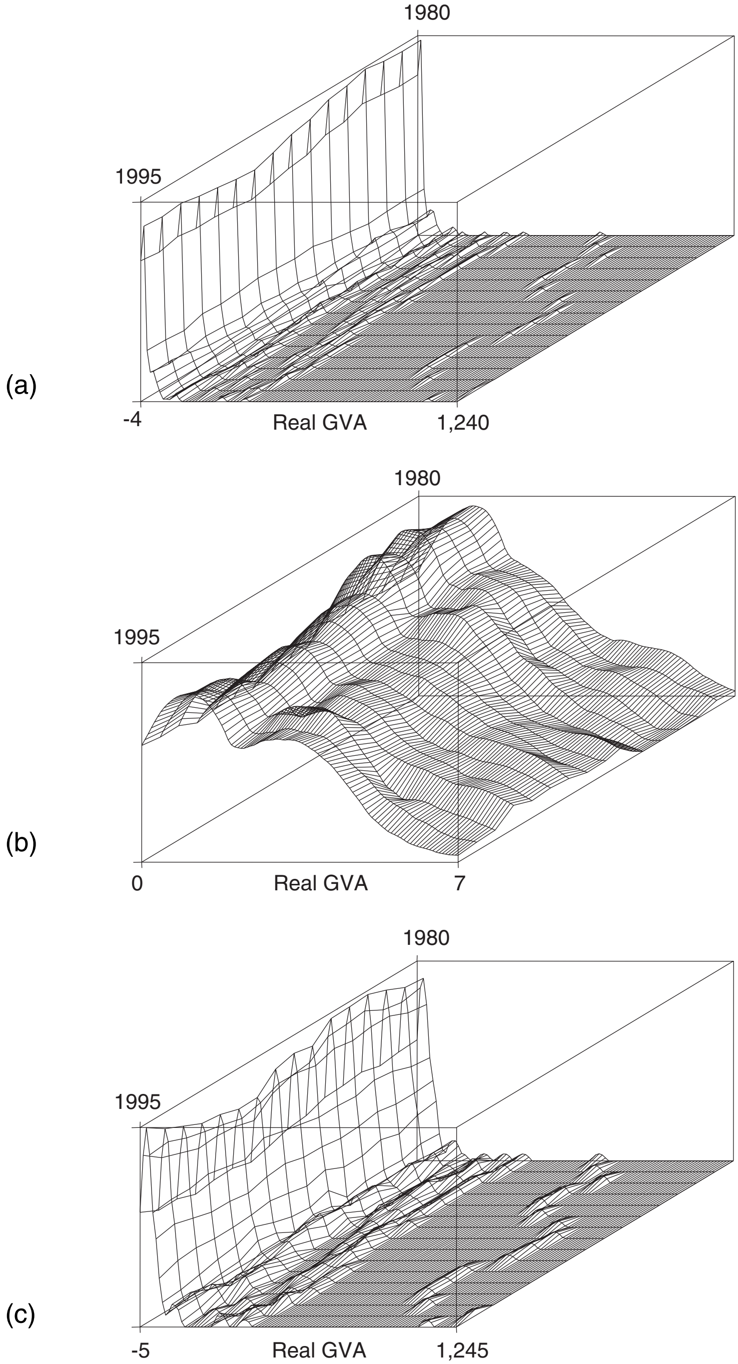


Figure 8.3a–c Kernel density estimation for population-weighted RGVA in European regions, 1980–92. (a) Full, (b) below- and (c) above-median.

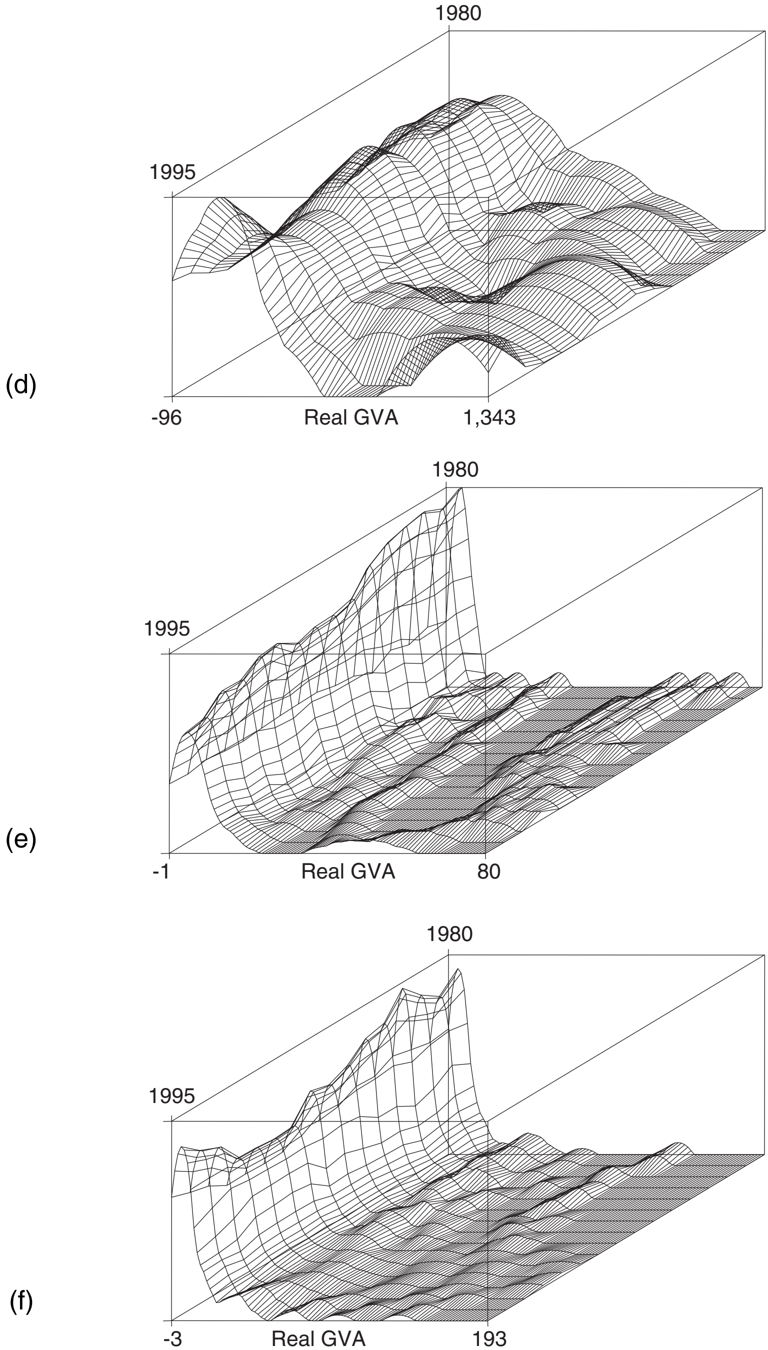


Figure 8.3d–f Kernel density estimation for population-weighted RGVA in European regions, 1980–92. (d) Centre, (e) adjoining, and (f) periphery.

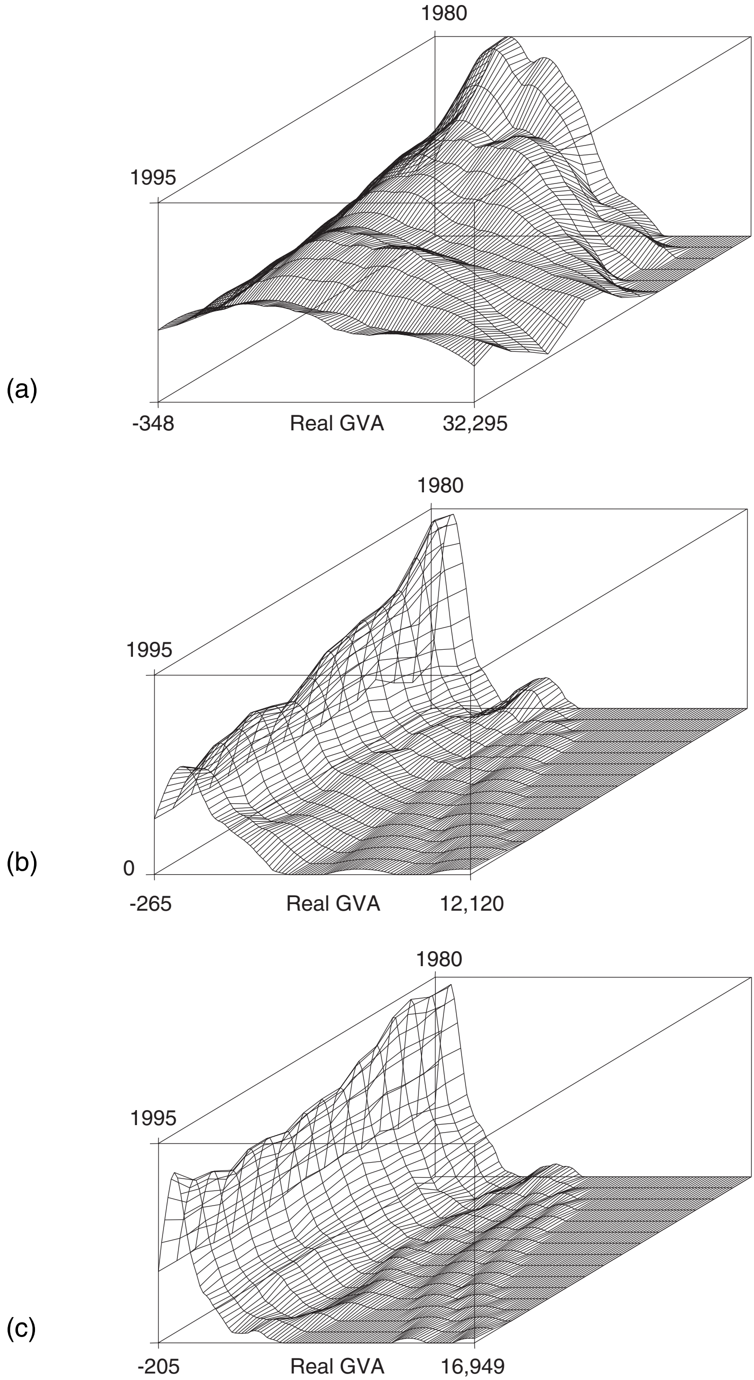


Figure 8.4a–c Centres, adjoining regions and the periphery, 1980–95.

5.3.1 Borders

To capture the effect of regulation, we test whether the subset of regions separated by a border behaves differently from the subset of regions that are not separated by a border. Therefore, we calculated the correlations between all border regions (24 regions with 27 borders). The correlations between the border regions are shown in Figure 8.5 and these are generally positive, as expected. Striking is the difference in correlation between The Netherlands, Germany and Luxembourg, where correlations are generally high, and Spain, Portugal and Italy, where correlations are low or even negative.[16] This implies that the northern European markets are more integrated than the southern European markets.

Do the computed correlations differ from the sample as a whole? The average correlation between all possible pairs in the sample is 0.25 while the average of the shown correlations is 0.39. This difference, however, is an overstatement since in the overall sample both regions that are close and further away are lumped together. We will map the regions more precisely shortly and then compute comparable correlation averages.

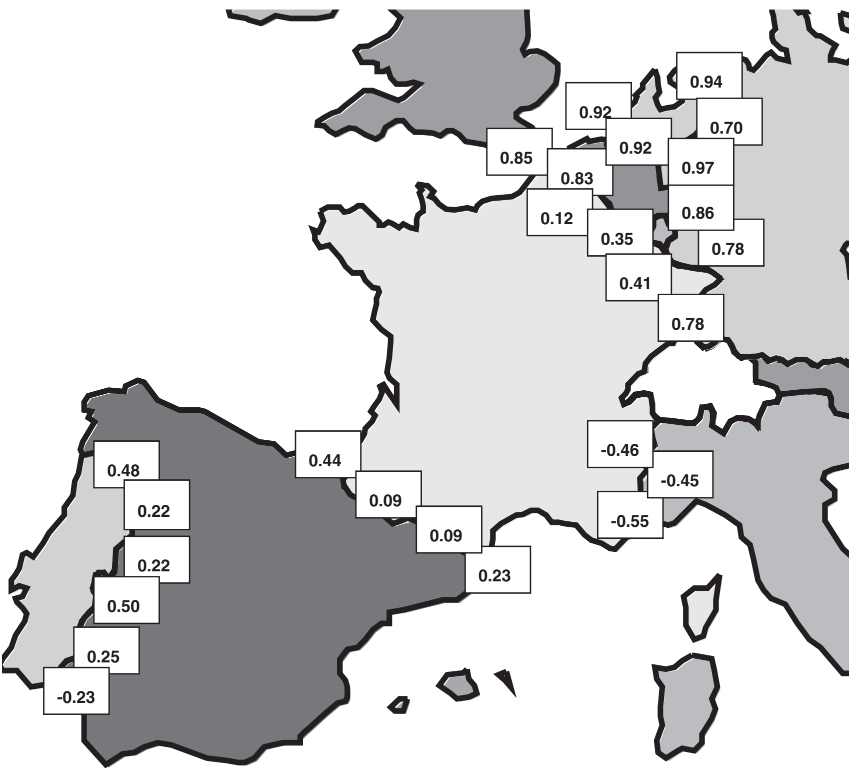


Figure 8.5 Correlation of Real GVA across borders, 1980–95.

Another interesting question is whether correlations have shown changes through time. To investigate this we plot a moving average (lag 8) of the correlations of all regions, and that of just border regions in Figure 8.6.[17]

The interesting feature of this graph is that there is quite some movement in the correlations between the regions. Moreover, correlation between non-border regions tends to be higher, probably because within-country correlations, which tend to be higher, are included in this set. Presently we are still at a loss for an explanation for the dynamics in this graph.

5.3.2 One market

As we investigate data over the period 1980–95, the completion of the EC internal market in 1992 could be another factor that has influenced the development over time of the distribution of GVA across regions. An important regulatory change has been the implementation of the Second Banking Directive into national law, which had to be completed by the beginning of 1993 at the latest. As the exact moment of implementation varies across countries (for example, in Germany 1992 and in The Netherlands 1993, see Lang, 2001), one clear 'breakpoint' cannot be identified. We considered 1992 as breakpoint and estimated the distribution before and after 1992. The dynamics in these graphs are rather subdued and at first sight, no significant change seems to occur. A more detailed inspection, however, shows that the behaviour of the data before 1992 seems to differ from that after 1992. Before 1992, the top of the distribution changes over

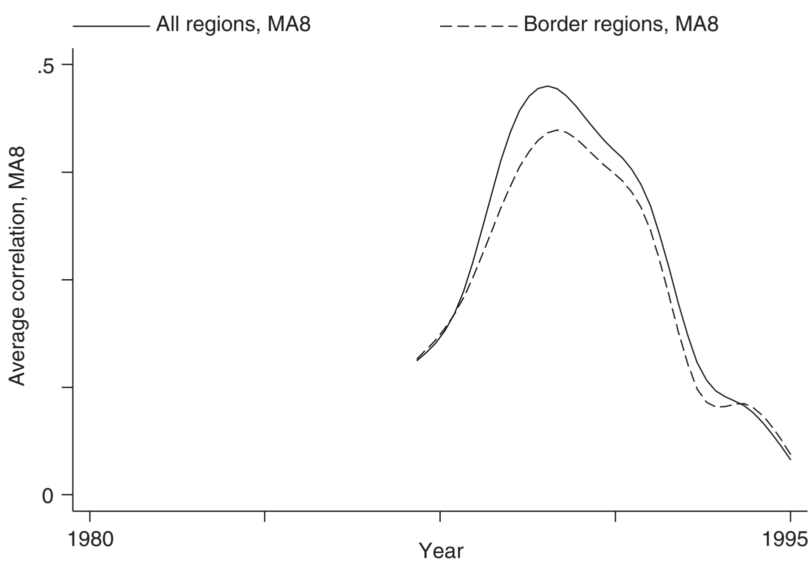


Figure 8.6 Moving average of correlations (lag 8).

time, whereas after 1992, it shows a stable pattern. Presumably, any changes due to the implementation of the One Market programme had already taken place. Brülhart (2001) finds similar results for manufacturing industries.

6 Employment

Our second measure of financial sector activity is employment in the financial sector. As in the case of the RGVA framework, we consider the development over time of employment in the financial sector. The period considered, however (1980–92), is shorter due to data limitations (for a detailed data description see Appendix 8A, Table 8A.2). Again we estimate the distribution of the data in each year in the period 1980–92. Figure 8.7 shows the graphical representation, where Figure 8.7a represents the full sample and Figure 8.7b and c show the distribution of the below- and above-median subsample. The overall distribution seems very stable. Looking at the below- and above-median subsets respectively, however, we see a development towards a distribution that is slightly more spread out. Compared with the developments in the real gross value added variable, the movements in employment seem rather modest.

6.1 The home market

Similarly to the RGVA variable, we examined whether the size of the home market, proxied by population, influences the distribution. In Figure 8.8 the estimates for the full sample and the below- and above-median sample are shown consecutively. Again, similarly to the earlier case, the effect of the home market seems to intensify the centrality of the distributions.

6.2 Centre and periphery

Does the centre–periphery chasm show up in the employment data? Whereas we saw some – minor – differences in the developments of the distribution of the various subsets, the distributions of the subsets shown in Figure 8.9 are rather similar. We will not show the equivalent population-weighted graphs because the developments are remarkably similar.

6.3 Regulation

Since our data on employment end in 1992 we would be hard pressed to investigate the employment effect of the One Market policies. Moreover, following the discussion in Section 5.3.2, above, it could be argued that 1992 would be too early a date to expect any changes in employment patterns due to One Market or the Second Banking Directive/Basel Capital Accord implementation.

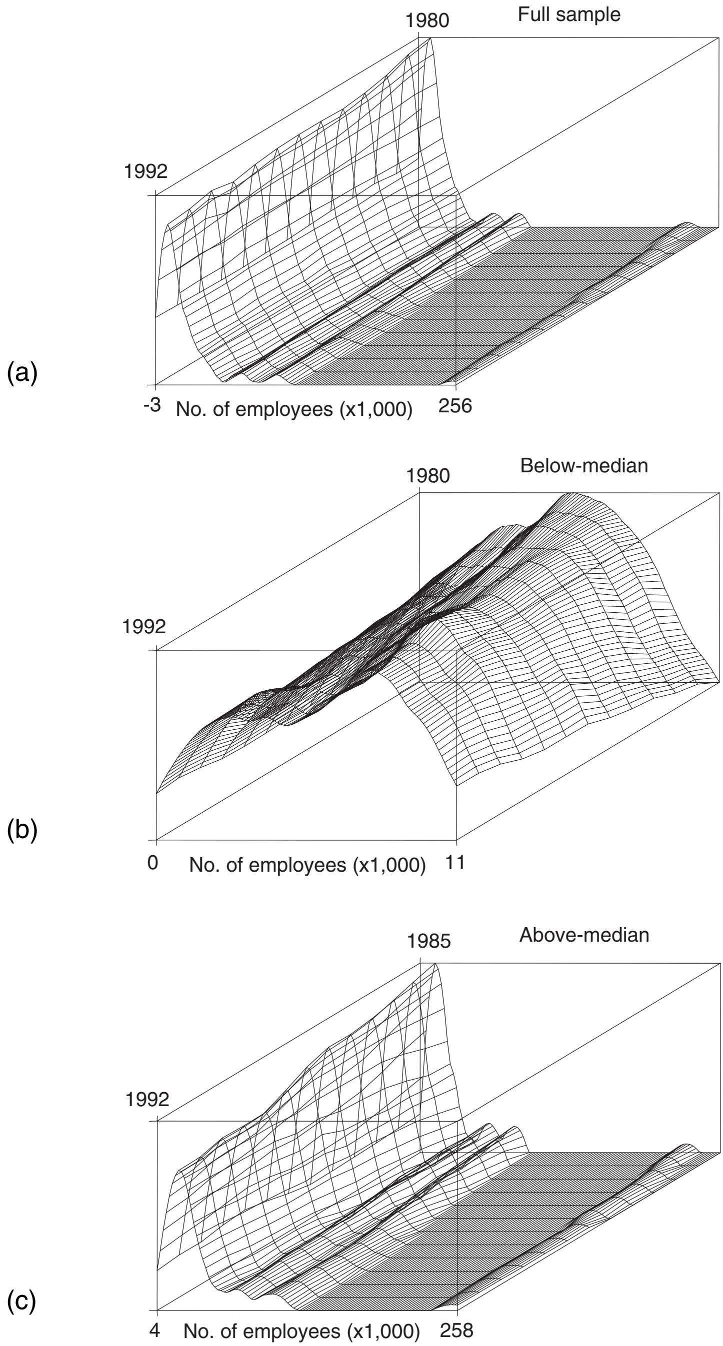


Figure 8.7a–c Kernel density estimation for employment in European regions, 1980–92. (a) Full, (b) below- and (c) above-median.

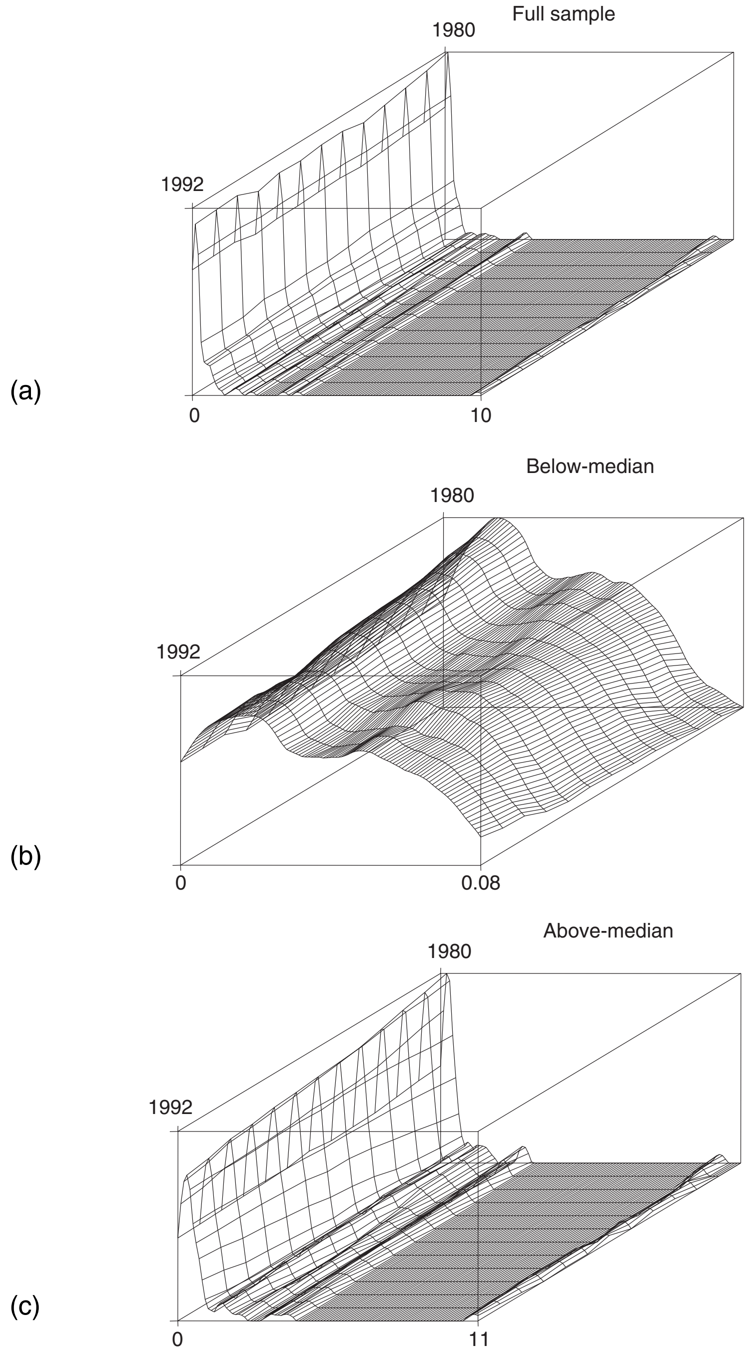
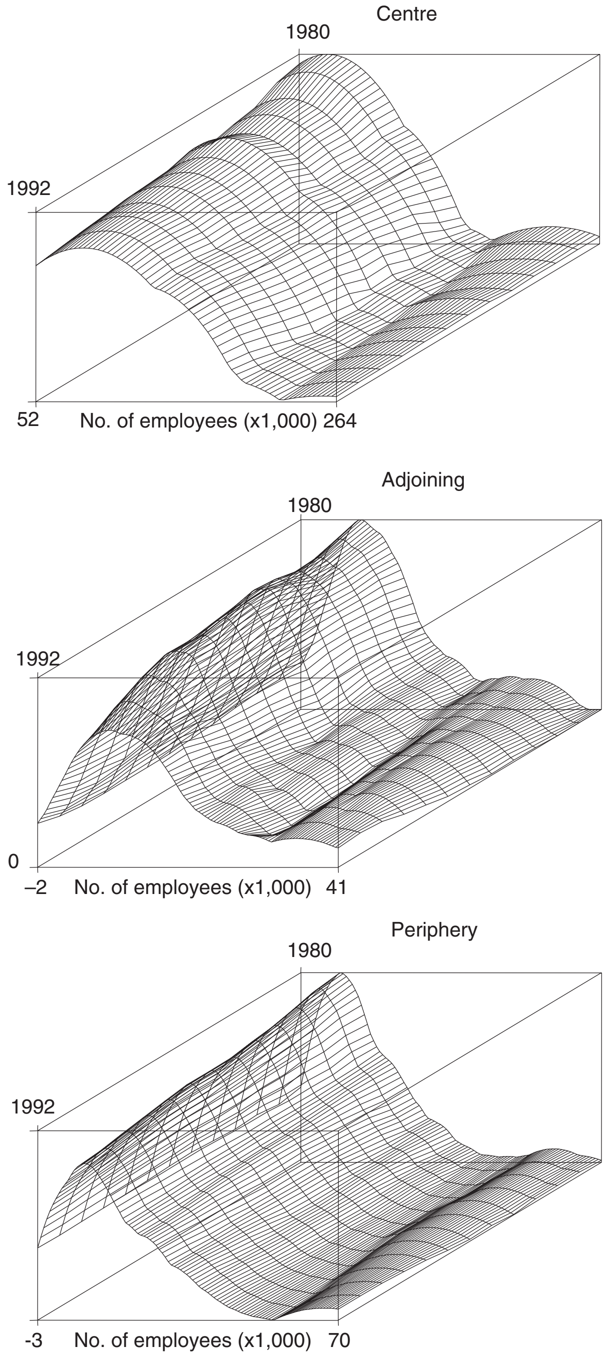


Figure 8.8a–c Kernel density estimation for population-weighted employment in European regions, 1980–92. (a) Full, (b) below- and (c) above-median.



Figures 8.9a–c Employment in centre, adjoining and peripheral regions, 1980–92. (a) Centre, (b) adjoining and (c) periphery.

6.3.1 Borders

As discussed previously, the existence of borders might very well influence the location of financial services. Moreover, different borders might have different effects. To examine this effect we have, similarly to the RGVA analysis, computed the correlation across all available border regions. Compared to the value added data the correlation across borders is generally higher. This conclusion gains in strength after a sensitivity analysis of various coefficients. Especially the two values found for the correlation between the south of Spain and the south of Portugal (–0.02 and –0.01) are sensitive to the length of the sample period because the underlying variable shows extremely low variance. The higher correlations might indicate that co-movement of 'financial sector activity' is indeed underestimated by the estimates shown in Figure 8.10.[18]

7 Efficiency

After looking at the development over time of value added and employment individually, we can also look at the combined picture. This would give us more

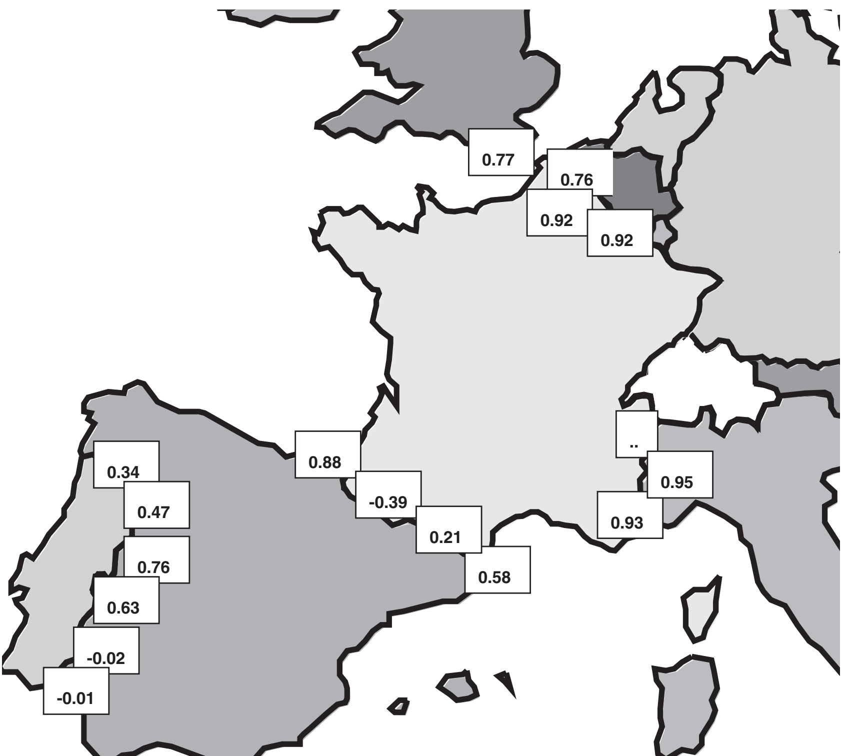


Figure 8.10 Correlation of employment across borders, 1980–92.

insight into the development of efficiency in the financial sector, since output per employee for the sector is an often used proxy (notwithstanding the well-known problems with measuring productivity in services).

The variable of interest, EFFI, is simply RGVA divided by employment. Estimates of the densities over time are shown in Figure 8.11. This variable is less

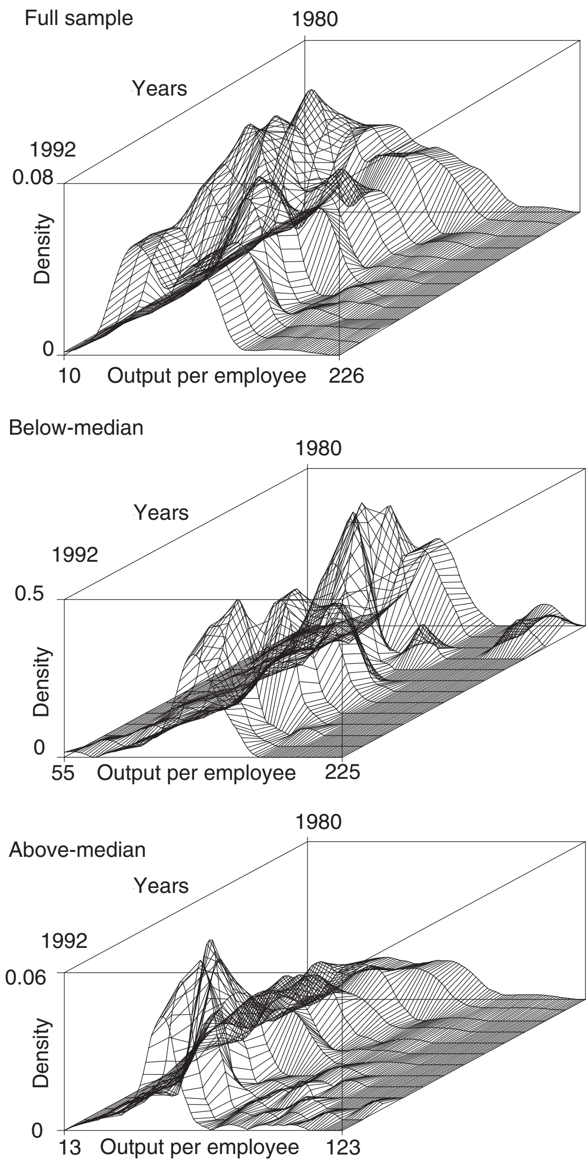


Figure 8.11 Kernel density estimation for efficiency in European regions, 1980–92.

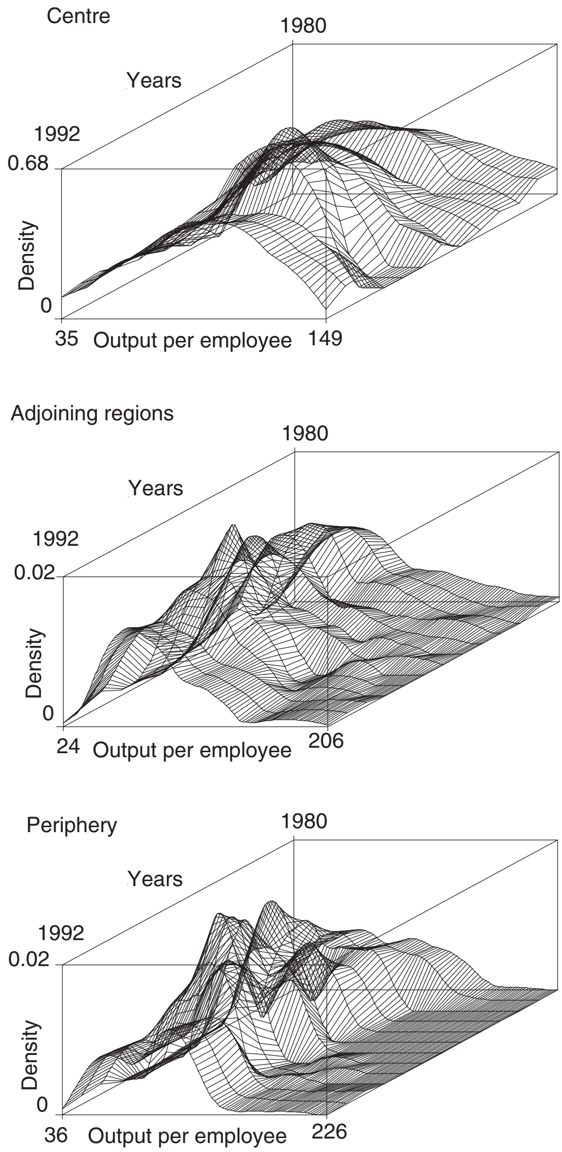


Figure 8.12 Efficiency in centre, adjoining and peripheral regions, 1980–92.

skewed but seems to exhibit double peakedness. The latter feature does not become clear from the figures in Figure 8.12 but does clearly show up on single-period density estimates. So far, a good explanation for it is lacking. Especially in the more efficient regions, our measure of efficiency seems to have become more concentrated.

We would expect really strong differences between the centre, the adjoining and the peripheral regions, especially in efficiency. Although we do see some movement in the middle of the sample period the distributions in the three subsets seem to have returned to their previous location (Figure 8.12).

8 Conclusions

In this paper, we have investigated whether technology has reduced the importance of geography in the financial sector, leading to a changed distribution of financial activities across space. After reviewing the centripetal and centrifugal forces identified in the literature, we approached this question empirically and analysed the development of the distribution of gross value added and employment for various European countries and regions.

We recognize that technology is just one of the numerous factors that influence the geographical pattern of financial activity. Because appropriate data are not available, we have not analysed if and how technology has changed the importance of the other forces, so that a good explanation of the final outcome is still lacking. From our analysis we conclude, however, that technology by itself has not (yet) been so powerful that significant changes in the estimated distributions can be observed. Alternatively one could conclude that either there has been insufficient change in the use of technology or all other factors combined exactly counter the effects of changed use of technology.

The distribution of financial activity, measured by GVA, has not moved significantly over the horizon 1980–95. A relocation of activities from centre- to low-cost areas, although technically feasible, cannot be observed. Also data on employment show no significant changes. The consideration of the effects of regulation and the completion of the internal market do not indicate a change in location either. With regard to efficiency there does seem to have been some movement, but at the end of the sample the distributions have taken up their old shape again. These results indicate that technology has not (yet) led to a changed spatial dispersion of financial activity. All in all, we find no evidence for the hypothesis that technology has eliminated the importance of geography completely as predicted by the 'Geography Doesn't Matter' hypothesis.

Concerning regulation, we tested whether regions separated by a border behave differently from those not separated by a border, and noted that the correlation of RGVA between border regions in the northern regions was rather high, whereas in southern regions correlations were low or even negative. Moreover, looking at the development of correlation through time we noted that border regions did not start to behave differently in the end. In addition we noted that there seems to be some movement in the levels of correlation between the regions.

A criticism of our analysis with some merit is that our data pre-date the coming of age of the Internet and that the Internet will make completely new ways of production possible. We do not disagree, but propose that the Internet will have an impact that is at least as strong as the impact of the other recent innovations in communication technology. Since recent data on Internet applications are not yet available in sufficient quantity and quality, our data are the next best thing.

A number of issues have not been resolved. Among these is the issue of the appropriate weight to use in order to take into account the home market effect. Possibly the size of other sectors in a region or the purchasing power of the population would be better proxies. Another issue is the computation of within-country correlations. These require additional work. Finally the underlying data on, especially, employment could be improved, both in quantity as in quality.

Appendix 8A

Table 8A.1 Description of Eurostat data for gross value added at market prices (GVA mp) or at factor cost (GVA fc), per country, 1980–95

Country	*Geographical level*	*Number of regions*	*GVA*	*Comments*
Belgium	NUTS 2	11	mp	
Denmark	National	1	mp	1995 missing
Germany (West, excl. Berlin)	NUTS 1	10	mp	NACE 6 data for NUTS 1 (Länder) available until 1994
Greece	National	1	mp	Some years in the 1980s for few branches missing
Spain	NUTS II	18	mp	1995 missing
France	NUTS II	22	mp	1995 missing; several years for Corse missing
Ireland	National	1	mp	Some years missing
Italy	NUTS II	20	fc	
Luxembourg	National	1	mp	Some years missing
Netherlands	NUTS II	12	fc	1994 and 1995 missing; some years in the 1980s for some regions missing
Austria	National	1	mp	
Portugal	NUTS II	7	mp	
Finland	National	1	mp	
Sweden	National	1	mp	
United Kingdom	NUTS I	11	mp: 1980–3 fc: 1984–95	
Total		119		

Source: Eurostat Regio database.
The Eurostat sectoral classification 'NACE 17'contains regional data of 17 branches and 5 groups of services. Category B69 provides data on GVA added by credit and insurance institutions. The difference between GVA at market prices and GVA at factor cost, being that the first includes indirect taxes and subsidies and the latter does not, is ignored.

Table 8A.2 Description of data on employment, per country, 1990–95

Country	*Geographical level*	*Number of regions*	*Comments*
Austria	National	1	ILO data
Belgium	NUTS II	11	
Denmark	National	1	
Finland	National	1	
France	NUTS II	22	
Germany			
Greece	National	1	
Ireland	National	1	ILO data
Italy	NUTS II	20	
Luxembourg	National	1	
Netherlands	National	1	National office for statistics (CBS). Employment is measured by the number of jobs (full- and part-time) in the financial sector and not by the number of employees, as for the other countries.
Portugal	NUTS II	7	
Sweden	National	1	ILO data
United Kingdom	National	1	ILO data

Source: Eurostat Regio database.

Notes

1 It should be noted that developing and transition economies may directly adopt the 'newer' technologies, i.e. the Internet and mobile communication, passing over preceding technologies ('leap-frogging'). This subject is analysed in detail by Stijn, Claessens *et al.* (2001).

2 Of course, different types of activities can be concentrated in different centres. A leading centre for forex-trading, for instance, does not necessarily imply a strong position in options-trading. Different activities are performed within different centres, which has led researchers, as for instance Scholtens (1992, p. 299), to conclude that in ranking financial centres, 'the geography of finance is not as strictly tripolar as generally assumed' (see also Tschoegl, 2000, p. 8).

3 We turn to the importance of path dependency later on in this paper.

4 See Keller (2001) for empirical evidence that R&D expenditures generate local spillover effects. Moreover, language seems to be an important determinant of the speed of technology diffusion.

5 Regulatory costs can be divided into direct and indirect costs. Direct costs are the costs of running the regulatory institution, if these costs are born evenly by market participants, and compliance costs. More important, however, are the indirect costs of regulation. Direct costs will lead to deviating prices which, in turn, lead to distortions in the markets. Supply might, for instance, be decreased, fewer products might be offered, and competition in financial markets might be reduced. Regulation might furthermore lead to regulatory arbitrage and distortions of a level playing field for banking.

6 We are grateful to David Llewellyn for pointing this out.
7 Other possible methods are one-way scatterplots, stem-and-leaf displays, and boxplots.
8 See Silverman (1986) for a detailed technical discussion of density estimation.
9 Due to still unsolved technical difficulties in drawing the graph, this is not clear from the graphs. The top-right of the box shown should show zero density.
10 The formal definition is given by: $G = 1 - 2 \int \varphi dF$, with $F(y)$ = the proportion of the population with GVA less than or equal to y = GVA. $\int$ implies that integration is performed over the entire range of y, that is, over [0, infinite] or, equivalently for F, over the range [0,1].
11 This measure can be computed as $c = \sqrt{\ } V/y^*$. Here, V is the variance and y^* the arithmetic mean RGVA.
12 In particular the relative mean deviation, standard deviation of logs, Mehran index, Piesch index, Kakwani index, Theil entropy index, and mean log deviation.
13 An alternative measure of the size of the home market might be income or disposable income.
14 A centre is defined as a country's region that makes the largest contribution to national GVA. Countries with only a single region, such as for instance Austria, are excluded, leaving a sample of seven countries (total number of observations: 1,648). These are The Netherlands, Belgium, Germany, Italy, the UK, Spain and France. The regions contain respectively the cities of Amsterdam, Brussels, Düsseldorf, Lombardia, London, Madrid and Paris. Adjoining regions are those regions next to a centre region. All other regions are called peripheral.
15 Differences between measured GVA could of course also be attributed to differences in accounting practices and definitions used.
16 Possibly the correlations are somewhat distorted due to the conversion into a common denominator: the euro. Presently, all values are first transformed into real values and then converted into euro. A more precise procedure for this particular question would be to transform the values at the contemporaneous exchange rate against the euro (or equivalent, possibly the German Mark).
17 Ideally we would like to compare the correlations between adjacent border and non-border regions. At present we have yet to classify the regions in this way. Therefore we take the shortcut of comparing adjacent border regions with the correlations between all regions. The latter is then probably an overstatement of the relevant correlation because regions further apart are likely to exhibit lower correlation.
18 Underlining the need to consider the effect of the exchange rate as already noted in 15 above.

References

Ades, A.F. and Glaeser, E. (1994) 'Trade and circuses: Explaining urban giants'. NBER Working Paper 4715.

Arthur, W.B. (1994) Preface to 'Increasing Returns and Path Dependence in the Economy', in W.B. Arthur (ed.) *Economics, Cognition and Society series*. Ann Arbor: University of Michigan Press: xi–xx.

Begg, I. (1991) 'The spatial impact of completion of the EC internal market for financial services', *Regional Studies* 26(4): 333–47.

Booz-Allen and Hamilton (1998) *Developments in Banking*, London: Booz-Allen and Hamilton.

Brülhart, M. (2001) 'Evolving geographical concentration of European manufacturing industries', *Weltwirtschaftliches Archiv* 137(2): 215–43.

Bryant, R.C. (1987). *International Financial Intermediation, Studies in International Economics*. Washington D.C.: Brookings Institution.

Buch, C.M. (2000) 'Why do banks go abroad? Evidence from German data', *Financial Markets, Institutions & Instruments* 9(1): 33–67.

Campayne, P. (1992) 'The impact of multinational banks on international financial centres', in M. Casson (ed.) *International Business and Global Integration*. Reading: Macmillan, pp. 121–62.

Claessens, S., Glaessner, T. and Klingebiel, D. (2001) 'E-finance in emerging markets: Is leapfrogging possible?', World Bank Financial Sector Discussion Paper 7.

Coase, R.H. (1937) *The Nature of the Firm*. Economica.

Cowell, F.A. (1995) *Measuring Inequality, LSE Handbook in Economics*. London: Prentice Hall.

Dufey, G., and Giddy, I. (1978) *The International Money Market.* New York: Prentice Hall.

Fujita, M., and Thisse, J.F. (1996) 'Economics of agglomeration', CEPR Discussion Paper Series.

Gaspar, J. and Glaeser, E.L. (1996) 'Information technology and the future of cities', NBER Working Paper 5562.

Gehrig, T. (1998) 'Cities and the geography of financial centres', CEPR Discussion Paper Series 1894.

Hallet, M. (2000) 'Regional specialisation and concentration in the EU', European Commission Economic Papers 141.

Isard, W. (1960) *Methods of Regional Analysis*. Cambridge, MA: MIT Press.

Jones, G. (1992) 'International Financial Centres in Asia, the Middle East and Australia: A historical perspective', in Y. Cassis (ed.), *Finance and Financiers in European History, 1880–1960*. Cambridge: Cambridge University Press.

Keller, W. (2001) 'Geographic localization of international technology diffusion', CEPR Discussion Paper Series 2706.

Kim, S. (1990) 'Labour heterogeneity, wage bargaining, and agglomeration economics'. *Journal of Urban Economics*: pp. 160–77.

Kindleberger, C.P. (1974) 'The formation of financial centers: A study in comparative economic history', *Princeton Studies in International Finance* 36.

Kindleberger, C.P. (1983) 'International banks as leaders or followers of international business', *Journal of Banking and Finance* 108: 583–95.

Krugman, P. (1991) 'History and industry location: The case of the manufacturing belt', *American Economic Review* 81(2): 80–3.

Lang, G. (2001) 'The impact of SMP and EMU on German banking', Working Paper, University of Augsburg 192.

Marshall, A. (1920) *Principles of Economics*. London: Macmillan.

Martin, R. (1999) 'The new geographical turn in economics: Some critical reflections', *Cambridge Journal of Economics* 23(1): 65–91.

McKillop, D.G. and Hutchinson, R.W. (1991) 'Financial intermediaries and financial markets: A UK perspective', *Regional Studies* 25(6): 543–54.

Nadler, M., Heller, S. and Shipman, S. (1955) *The Money Market and its Institutions*. New York: Ronald Press Company.

Newman, P. and Kenworthy, J. (1989) *Cities and Automobile Dependence*. Gower: Aldershot.

O'Brien, R. (1992) *Global Financial Integration: The End of Geography*. London: Council on Foreign Relations Press, Royal Institute of International Affairs.

Pagano, M. (1989) 'Trading volume and asset liquidity', *Quarterly Journal of Economics* 104(2): 255–74.

Porter, M.E. (1998) 'Clusters and the new economics of competition', *Harvard Business Review* 76(6): 77–90.

Porteus, D.J. (1996) *The Geography of Finance*. Brookfield: Aldershot.

Rauch, J.F. (1993) 'Productivity gains from geographic concentration of human capital: Evidence from three cities', *Journal of Urban Economics* 34: 380–400.

Reed, H.C. (1981) *The Pre-eminence of International Financial Centres*. New York: Praeger.

Rosen, R.D. and Murray, R. (1997) 'Opening Doors: Access to the global market for financial sectors', in M. E. Grahan and A. Vourvoulas-Bush (eds), *The City and the World: New York's Global Future*. New York: Council on Foreign Relations.

Scholtens, L.J.R. (1992) 'Centralization in international financial intermediation: Theory, practice, and evidence for the European Community', *Banca Nazionale del Lavoro Quarterly Review* 182: 255–304.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Smith, D.M. (1981) *Industrial Location: An Economic Geographical Analysis*. London: John Wiley & Sons.

Stuart, C. (1975) *Search and the Organization of Market Places*. Malmo: Lund Economic Series.

Thrift, N. and Leyshon, A. (1988) 'The gambling propensity: banks, developing country debt exposures and the new international financial system', *Geoforum* 19: 55–69.

Towey, R.F. (1974) 'Money creation and the theory of the banking firm', *Journal of Finance* 29(1): 57–72.

Tschoegl, A.E. (2000) 'International banking centers, geography, and foreign banks', *Financial Markets, Institutions & Instruments* 9(1):1–32.

Walter, I. (1998) 'Globalization of markets and financial centre competition', in R. J. Langhammer (ed.), *Challenges for Highly Developed Countries in the Global Economy*. Kiel: Institut für Weltwirtschaft.

Weber, A. (1929) *Theory of the Location of Industries*. Chicago: University of Chicago Press.

Part III

Contributions on technology and financial markets

9 Competition and consolidation in the European exchange industry*

Olivier Lefebvre

This contribution focuses on the reason why consolidation is needed in the European exchange industry, and on the way this consolidation is likely to take place. The first section focuses on the competitive environment of this industry and the key drivers to consolidation. The second section highlights the differences between two approaches to consolidation and rationalization: the old (conservative?) view versus the Euronext view. Finally some comments are formulated on the European Regulatory Framework, which is of major importance for the future shape of this industry.

1 Competition and need for consolidation

Competition between exchanges

It is often said that with the computerization and the internationalization of the activity, competition between exchanges became tough. This might be true, but it is important to understand the nature of this competition. On the primary market, exchanges are supposed to compete for listings. In fact, this competition only concerns a marginal number of cases. Most of the IPOs in Europe take place on the domestic market, even for big global-sized companies. This strong national bias can be explained by two reasons. The first one is regulation. Given regulatory discrepancies (IPO regulation, ongoing transparency obligations, accounting standards) it remains more natural for a company to get listed in its jurisdiction of registration. The second reason has more economic ground: it is the home market factor which remains important for the equity market. This is particularly, though not exclusively, true for small and mid caps. Issuers are indeed keen to tap the domestic investor pool where they benefit from a natural notoriety, in a sort of informational hinterland.

On the secondary market, exchanges are expected to compete for liquidity, and try to attract to their order book the flow of orders that normally goes to their competitors. Evidence shows that the volume traded on non-domestic shares on the order book of the various European exchanges, is extremely limited. The explanation is simple: according to an empirical study by Elkins/McSherry, the

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

direct exchange costs represent only a minor part of the total transaction cost for the investor. A much larger part is due to the price impact of the transaction which is linked to the depth of the market. This creates a barrier to entry. Indeed, once an exchange concentrates most of the buy and sell orders on a specific security, it is difficult for another exchange (even though much more competitive) to attract the order flow. In other words: liquidity is sticky.

For all these reasons, direct competition between exchanges is rather limited, and relative performances of European stock exchanges (in terms of market shares) are for 90 per cent explained by the performance of their national blue chips, not by their direct competitive ability. European exchanges do not compete on a true pan-European scale, but rather at the margin. However, this does not mean there is no other type of competitive pressure. Indeed, big users who are members of many different exchanges, do compare the quality and prices of services offered by the various exchanges. As a consequence, the big users are putting a lot of pressure on them.

Pressure from new entrants

The Alternative Trading Systems (ATS) and electronic communication networks (ECNs) have a significant market share in the USA: 35 per cent of NASDAQ and 12 per cent of NYSE. This can probably be explained by the traditional operating techniques of the established marketplaces, which did not offer automated matching, anonymous trading, etc., offering a window of opportunity for ATS. So far, the ATS have had no success in Europe because the existing organized markets are working with a strongly developed screen infrastructure based on the central order book market model.

Another phenomenon important to competition is the growing significance of the so-called 'in-house' matching of securities transactions among the customers of large banks. This potentially is the most serious threat for exchanges. It is stimulated by the geographical fragmentation of European exchanges. In-house matching raises questions about transparency, best execution and the possibility of conflict of interest. It might also fragment liquidity. We will come back to this issue in the final section on regulation.

The two main drivers for consolidation

With more than 40 independent operators in Europe, the exchanges industry is highly fragmented. Two drivers should lead to a rather rapid consolidation: the pressing demand of the market participants and the cost structure of the exchanges.

Market participants internationalize on both sides of the markets. The investors look at sectors rather than at national markets. Intermediaries consolidate on a cross-border basis. This forces market and backoffice operators to provide straight-through processing (STP) solutions on a cross-border basis. These STP solutions are expected to be at least as efficient as the national ones. The larger issuers, on their side, to a growing extent consider their 'domestic' capital market to reach beyond their national borders.

The other driving force is the cost structure of the highly IT-intensive exchanges themselves. Just like a network operator, an exchange has a significant infrastructure cost, trading and clearing systems, dissemination and access to many members, but very low marginal costs. The cost of listing one more company or connecting one more member is close to zero. With this cost structure, consolidation creates significant synergies, by replacing full IT costs with low marginal ones.

2 Two strategic approaches to European consolidation

Let me distinguish between two approaches. The first one, which I tend to call the 'old' view, consists in a national exchange claiming it is big enough to attract all the issuers and the members, bringing the liquidity into the national market and regulatory environment. This approach looks attractive because of its apparent simplicity: no need for migration or regulatory harmonization. This has been tried several times, without success so far; let us think of SEAQ international, EASDAQ, Tradepoint, and more recently Jiway, Virt-X or the new European blue chips segment on SETS. There are several difficulties, the first one being the already mentioned 'stickiness' of liquidity. Moreover this approach triggers resistance from the clients because it puts the adjustment costs on them (listed companies, members) and destroys the 'home market value' of European equity markets.

Euronext has developed a new approach to the world of exchanges, by applying a principle which is very common in any other kind of business: deliver services in the jurisdiction of the clients. By keeping the national franchises as regulated markets, while unifying the trading platform and the rulebook, it combines the existing pools of liquidity on a cross-border basis. This approach, although apparently more difficult, notably in terms of harmonization efforts, is more client-friendly and provides several advantages:

- It brings continuity to existing clients. Already listed firms and existing members of one of the exchanges will benefit from the new dimension of the market, without having to move to another jurisdiction.
- New clients can choose the entry point according to their preferences.
- It fully protects the positive 'proximity dimension' of the home market, while still providing system synergies and cost cutting, as well as international exposure and enhanced liquidity.

Let us now go into the practicalities of this Euronext approach.

On the regulatory side, members and listed companies choose their entry point under French, Belgian or Dutch law. The market is unified by one single rulebook, the so-called 'Euronext Rulebook One', which is applicable under the various jurisdictions, since it was approved jointly by the national regulators. The present dimension of Euronext with three jurisdictions is aimed at being enlarged to other ones without changing the basic rules (see Figure 9.1).

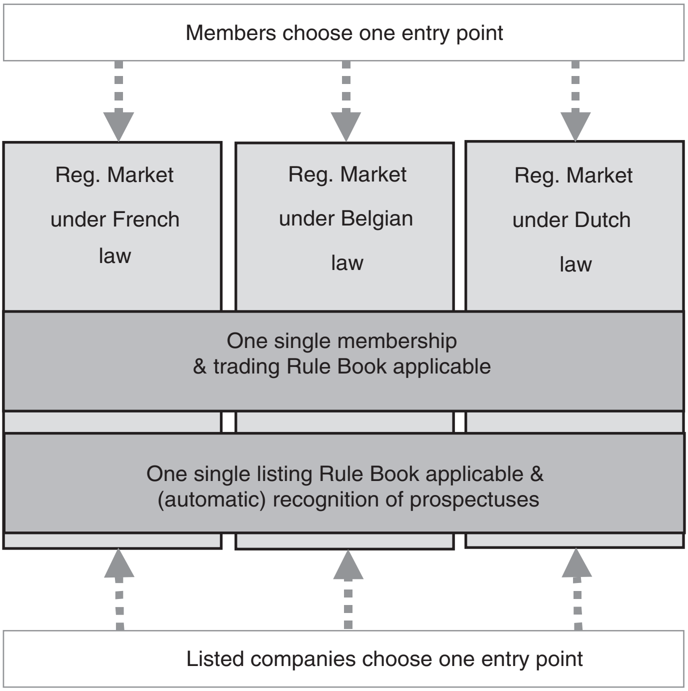


Figure 9.1 The 'Euronext' view.
Source: Euronext.

This approach allows for important synergies. A single unified platform for cash trading has been created. The Euronext market model version of NSC was launched in Paris in April 2001. Since May and October 2001, respectively, the Belgian and the Dutch markets are operated on that single platform, allowing the Brussels (NTS) and Amsterdam (TSA) systems to cease operations. A similar approach is being prepared for clearing and settlement (with Clearnet and Euroclear) and the same will apply to the derivatives market. The expected synergies on information technology are estimated at 52 million euro per annum, which represents 25 per cent of the total IT cost prior to the integration. The single listing rulebook and the automatic recognition of prospectuses are, at this stage, still to be completed.

The vertical integration, which is illustrated in Figure 9.2, is expected to create huge efficiencies for clients. However this vertical integration should not be seen as a 'silo'. On the contrary, it is designed to be open at every stage of the operational chain.

3 Regulatory environment in Europe

The existing European regulatory framework is clearly a big hurdle and a source of cost in the efforts to integrate the European capital market. Euronext clearly supports the efforts of the European institutions to improve this framework.

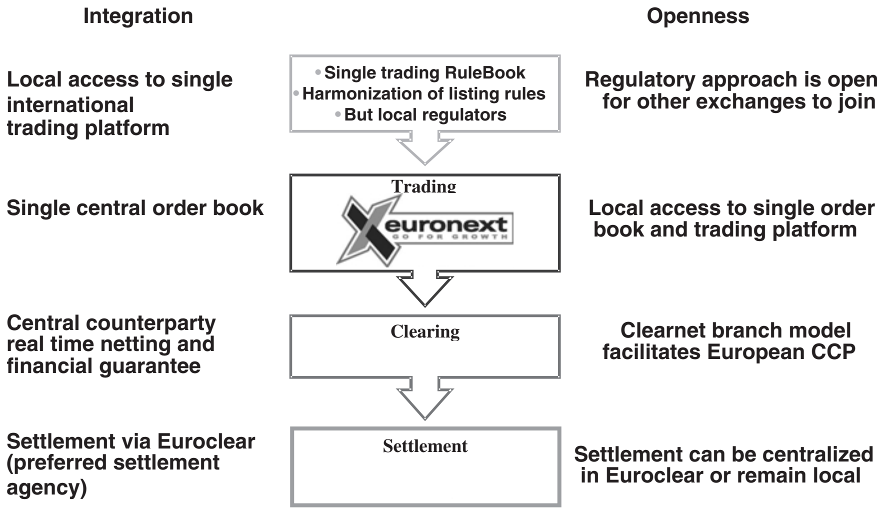


Figure 9.2 Integration and openness within the 'Euronext' view.
Source: Euronext.

However, we clearly fear that the current initiatives will fall short of reaching the following key objectives:

- exchanges should be allowed to operate on a truly pan-European scale;
- true direct competition between exchanges should exist. This means all regulated markets should be allowed to deliver their listing and trading services under the 15 EU jurisdictions.
- a level playing field should be established between regulated markets and alternative services providing orders matching services, ATS or 'in-house matching'.

In order to reach these objectives, the new European directives should meet the following principles:

- Issuers' legal obligations (notably, regarding prospectuses and transparency ongoing obligations) should be harmonized and controlled by an administrative (possibly home) authority independently from the choice of the market for listing. In this case, issuers would choose the market for listing on its own merit, without regulatory or national bias.
- Members' legal obligations (notably, conduct of business rules) should be harmonized and controlled by an administrative (possibly home) authority independently from the choice of the market on which they trade. Even if less critical than the previous one, it would enhance fair competition between exchanges.
- The definition and recognition procedure of regulated markets should be harmonized and so should be the principle that all market-specific rules should

be contractual (subject to approval by the competent authority). Therefore, contracts between exchanges and their clients (issuers and members) could be governed by the laws of any of the EU Member States, at the choice of the parties. The control and enforcement of these rules should be performed by the exchanges, based on the contractual relationship (i.e. private law principles).

This consistent framework would not only meet the objectives mentioned above, but would also deliver at least two other advantages. First, since market-specific rules would be contractual, this would allow for a fairly good degree of flexibility in the differentiation of market models (order driven versus quote driven, etc.) or segmentation (main list, new markets, etc.). This is crucial for the dynamic of the European marketplace. Second, statutory harmonization would be limited to 'general interest' rules and would be of general application for all issuers, intermediaries and markets, thus creating a level playing field in the exchanges industry.

The need for a level playing field requires that the regulatory framework establishes rules for alternative order matching mechanisms, including 'in-house matching' or 'internalization' by large intermediaries. In other words, the revised ISD should foresee regulation by function rather than by nature of institution. All order matching services should be regulated similarly. Indeed, if 'in-house matching' does not follow the same rules as for orders executed on regulated markets, a number of perverse effects will ensue:

- higher spreads on the marketplace;
- withdrawal of orders from price discovery mechanism will reduce informational content of the prices;
- no check of 'best execution' possible;
- possible conflict of interest due to the multicapacity of intermediaries.

4 Conclusion

Consolidation of the European exchange industry, from trading to settlement, is needed and will provide significant cost reduction. This will contribute enhancing competitiveness of the European financial industry and a reduction of the cost of capital; it will contribute to a more efficient European economy.

With this goal in mind, consolidation requires bringing national markets together rather than trying to move them to one particular national jurisdiction. Denying the role of the home equity markets is destroying value, especially for the small and mid caps. Technology allows exchanges to serve clients where they are offering them the dimension of a European market. Regulation should not slow down the pace of cross-border consolidation, but it should help exchanges to go in that direction.

10 Alternative Trading Systems and liquidity*

Hans Degryse and Mark Van Achter

1 Introduction

Traditional exchanges face enormous challenges. Technology, deregulation and investor needs are driving forces reshaping the trading landscape throughout the world. Technological progress not only allows direct access at traditional exchanges. It also enables the creation of new marketplaces, called 'Alternative Trading Systems' (ATS). A general definition of ATS is 'a trading mechanism developed independently of the established marketplaces and designed to match buyers and sellers on an agency basis' (Salomon Smith Barney, 2001). The purpose of this paper is to review the importance of ATS and their impact on the liquidity of traditional marketplaces.

ATS have gained success in the USA after the introduction of new Order Handling Rules in 1996 followed by 'Regulation ATS' in 1998. The latter regulatory measure mainly improves the linkages between traditional markets and ATS by requiring ATS to become self-regulatory organization members. We will show that in the USA, ATS have been particularly successful in attracting trade in the NASDAQ dealer market whereas they are less successful in competition with the NYSE. In Europe, traditional trading marketplaces automated earlier than in the USA. Moreover, continental European exchanges are typically organized as auction systems implying an agency nature of trading. The liquidity externality then makes it more difficult for ATS to develop a successful business model in Europe.

The market microstructure of ATS and traditional marketplaces is a major determinant of their future success. This literature is mainly concerned with the process by which investors' latent demands are ultimately translated into transactions. Given the different driving forces transforming the trading landscape, market microstructure helps in judging the relative merits of the different designs of the ATS. It also helps in making projections on their impact on liquidity in the traditional marketplaces.

The remainder of this paper is organized as follows. Section II starts by offering evidence on the impact of technology/automation on trading costs. Section III develops a typology of traditional exchanges and ATS. The fourth section deals with the relation between ATS and liquidity. The final section concludes.

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

2 Technology and trading costs

Domowitz (2001) argues that the automation of traditional financial markets plays an important role in the evolution of the industrial organization of the trading services industry. He propounds that markets are firms with network externalities related to liquidity. Although the link between liquidity and trading costs is well known, he is the first to investigate the connection between the automation of market structure and trading costs.

Table 10.1 provides an overview of the trading costs for the period 1996–99. The execution costs are based on data gathered by Elkins/McSherry and were published in *Institutional Investor.*[1]

Transaction costs are falling worldwide, illustrated by the decline of the average total trading cost from 73 basis points in 1996 to 61 basis points in 1999. Explanations include competition for order flow, shifts of trading strategies to accommodate liquidity differences, more institutional trading, and pressure from new trading systems and regulatory authorities.

Table 10.1 Total trading costs in 42 countries

Country	*1996 (basis points)*	*1997 (basis points)*	*1998 (basis points)*	*1999 (basis points)*
Argentina	93.1	59.7	48.7	62.7
Australia	57.0	53.8	47.0	54.2
Austria	40.3	39.9	54.1	42.7
Belgium	37.1	31.1	33.9	27.9
Brazil	63.1	53.7	46.6	47.1
Canada	63.0	51.6	43.9	39.9
Chile	115.4	60.4	47.0	131.0
Colombia	99.6	91.7	95.1	96.3
Czech republic	134.6	150.5	161.0	71.2
Denmark	35.7	45.4	43.4	41.1
Finland	41.9	42.3	44.0	40.7
France	29.9	26.7	26.6	24.9
Germany	39.3	33.3	27.6	28.7
Greece	64.4	66.9	63.6	87.3
Hong Kong	59.2	56.6	50.1	43.7
Hungary	145.2	163.7	102.3	71.6
India	85.8	65.0	64.8	128.7
Indonesia	108.5	92.4	95.5	84.8
Ireland	153.3	105.1	99.4	71.9
Italy	36.1	29.7	30.4	34.2
Japan buy	30.6	26.5	18.2	25.1
Japan sell	56.0	47.1	36.3	25.1

cont.

Luxembourg	75.5	73.0	70.0	102.3
Malaysia	87.3	87.8	90.8	90.7
Mexico	69.3	54.7	61.0	55.6
Netherlands	69.3	25.8	30.0	28.4
New Zealand	53.6	38.5	38.9	35.3
Norway	46.1	34.0	36.4	34.4
Peru	93.9	80.1	76.0	89.6
Philippines	114.9	107.5	105.0	109.0
Portugal	62.7	59.9	41.1	42.7
Singapore	71.9	76.6	84.9	64.9
South Africa	89.6	68.3	58.5	80.1
South Korea	228.9	200.1	97.8	78.9
Spain	47.1	34.9	43.0	42.3
Sweden	36.1	30.6	30.9	31.5
Switzerland	37.1	44.0	46.0	36.5
Taiwan	72.9	66.5	56.8	54.0
Thailand	93.8	87.2	75.5	82.6
Turkey	77.2	68.4	57.1	40.5
UK buy	73.7	75.1	71.0	71.1
UK sell	32.8	30.1	34.2	30.5
US NYSE	34.1	31.5	23.6	24.6
US OTC	51.9	39.0	29.9	33.3
Venezuela	113.4	158.4	144.7	195.8
Average	73.2	65.9	59.6	60.8

Source: *Institutional Investor.*

Domowitz (2001) investigates whether the adoption of an automated trading technology on traditional exchanges actually contributes to trading cost reductions. He sheds light on this issue at an international level (42 countries). In this section we first briefly summarize the methodology used in this study and its main results. Next, our own empirical results specifically for Europe will be presented.

2.1 International empirical evidence

Domowitz uses the Elkins/McSherry data on trading costs. This allows him to make a distinction between explicit and implicit costs. These are respectively related to development and operating costs[2] (i.e. fees and commissions), and to the dissemination of information on liquidity[3] (i.e. indirect trading costs like price impact costs, including the bid–ask spread). Evidently, the various cost components may be linked to each other. For example, minimizing price impact may imply incurring higher commissions. The link between automation and savings in explicit trading costs is quite obvious. Implicit trading costs, however, are at first sight not directly related to the automation of the market structure as liquidity is only created by the traders' presence on the system. But obviously, an automated market system indirectly affects liquidity as its design affects traders' incentives and capacities to monitor the market. Therefore, automation may shape the properties of transactions prices and market efficiency.

Next, Domowitz tests whether total trading costs and their components depend on the adoption of an automated trading technology by using regression-based techniques. As control variables, volatility, turnover and market capitalization are used. The regressions are performed on a quarterly, cross-sectional basis for all the countries in the dataset in the period from 1996:4 to 1998:3. Some of these countries do have exchange facilities that are largely automated with respect to execution while others do not.

The results for the international sample indicate that markets that are largely automated have average total trading costs that are, *ceteris paribus*, 33 to 46 basis points lower than those of their non-automated counterparts. Both types of cost components, explicit and implicit costs, hinge on automation. Explicit costs are between 23 and 32 basis points lower, whereas implicit costs are 10 to 18 basis points lower. Thus, on an international level the automated trading market microstructure does seem to have an effect on costs. This difference might be related to the higher floor development costs on non-automated exchanges.[4]

Domowitz (2001) discerns why the automation of markets permits the realization of implicit cost savings. By which means do electronic market systems allow traders to reduce price impact costs? One answer to this question is the presence of an electronic limit order book. Via this tool, which characterizes automated markets, traders can easily and instantly monitor certain liquidity characteristics (i.e. strategic liquidity management (see note 3)). This allows traders to execute their transactions when the market is rather liquid implying a reduction of transaction costs. Indeed, in reality the data indicate that traders do tend to use the electronic system to monitor liquidity and trade in a strategic way using this information. The price impact of realized trades is much smaller than that of trades executed under a naïve trading strategy that ignores monitoring of the book and stays almost constant along different trade sizes.

2.2 Empirical evidence for Europe

The analysis of automation in Europe on the basis of the Elkins/McSherry data is rendered somewhat trickier. The reason is that traditional European stock markets are technologically advanced and many exchanges were already electronic since 1996, which is the starting date of the Elkins/McSherry dataset. However, as can be seen from Table 10.2, there are some notable exceptions.

In order to compare our results with those of Domowitz (2001), we replicate his regression for the European countries only. More specifically, we estimate the following equation:

$$\begin{aligned} TradingCost_{it} = {} & \beta_0 + \beta_1 AutomationDummy_{it} + \beta_2 MarketCapitalization_{it} \\ & + \beta_3 Volatility_{it} + \beta_4 Turnover_{it} + \beta_5 YearDummies + \varepsilon_{it} \end{aligned}$$

Table 10.2 Electronic trading systems in Europe

European exchange	*Electronic since (year)*
Amsterdam	1994
Austria	1999
Borsa Italiana	1994
Brussels	1996
Copenhagen	1999
Deutsche Borse	1992
Finland	1997
London Stock Exchange	1997
Madrid	1989
Oslo	1999
Paris Bourse	1988/1994
Stockholm	1989
Switzerland	1996

Source: Internet and Salomon Smith Barney.

The results, which are displayed in Table 10.3,[5] show a significant negative coefficient for the electronic trading dummy. The interpretation is that in automated markets total trading costs are about eight basis points lower.[6] The savings in explicit trading costs are somewhat higher than those in implicit costs. Although the results should be interpreted carefully due to the low number of countries and the short time period, they confirm the negative coefficients obtained in Domowitz (2001) for 42 countries. However, the magnitude of our coefficients is substantially smaller than the conditional savings in international trading costs due to automation as reported in Domowitz (2001). In particular, he reported total costs savings of 33 to 46 basis points, explicit cost savings of 23 to 32 basis points, and implicit cost savings of 10 to 18 basis points. These are about five times larger than the results obtained for Europe only, which suggests that the impact of automation is less pronounced for this continent. Two explanations may drive these differences. First, it is possible that the automation dummy may only capture another

Table 10.3 Impact of electronic trading systems on trading costs in Europe

Dependent variable	*Total costs*	*Implicit costs*	*Explicit costs*
Electronic trading dummy	–7.82	–3.29	–4.52
	(2.72)	(1.78)	(1.59)
# country-years	52	52	52

Note: OLS estimates of the three proxies of trading costs (total, implicit and explicit costs) on an ‘electronic trading dummy’ and other control variables (market capitalization, volatility, turnover and year dummies). Standard errors are in parentheses.

step towards a fully electronic market. A second explanation is that the automation dummy in Domowitz may also capture agency trading or deregulatory effects. Agency trading is dominating in (continental) Europe even when automation was not yet in place.

3 Typology of traditional markets and Alternative Trading Systems (ATS)

The market microstructure literature typically distinguishes dealer markets and auction markets. Market makers are the only providers of liquidity in dealer markets. They are a counterparty in all transactions and quote two prices: the bid price, at which they are willing to buy securities, and the ask price, at which they will sell. The difference between those two prices is the market maker's spread. This spread hinges on the degree of asymmetric information between the dealer and informed traders, inventory costs and the remuneration for the service of providing immediacy (see Glosten and Milgrom, 1985; Ho and Stoll, 1981 and Demsetz, 1968, respectively). An example of a dealer market is NASDAQ. On auction markets, on the contrary, investors trade directly with each other or with the intervention of a broker-dealer acting as an agency trader only. All unexecuted orders are gathered in a limit order book. Market orders consume liquidity. Limit orders that do not execute immediately supply liquidity and could therefore be seen as free (short-lived) options against which market orders can be executed. Examples of auction markets are Euronext and the Toronto Stock Exchange. Other important characteristics are the degree of continuity of the exchanges, the degree of price discovery and the transparency (see Madhavan, 2000 for a review). Some only operate at certain points in time during the day whereas others are continuous.

There is a wide variety in Alternative Trading Systems (ATS). In referring to ATS we exclude the established marketplaces (traditional exchanges) as well as 'internal netting systems' (organized by individual intermediaries). A typical aspect of ATS concerns the fact that buyers and sellers meet on an agency basis.[7]

Within the ATS, we distinguish three groups of networks for which we will present a brief description of their typical features.[8]

A first important category is electronic communication networks (ECNs). Weston (2001) describes ECNs as 'electronic trading systems that allow investors to clear trades through an open limit order book. Rather than place orders with a specialist or dealer, traders on ECNs may anonymously submit orders and trade with each other directly'. A typical feature is that brokers on this communication network are acting on an agency basis only. ECNs allow traders to submit priced trades, i.e. limit orders. Therefore, ECNs have the potential to contribute to price discovery. Most ECNs guarantee pre- and post-trade anonymity.

A second category of ATS is external crossing networks. The SEC (2000) defines crossing networks as 'systems that cross multiple orders at a single price and that do not allow orders to be crossed or executed outside of the specified times'. Crossing systems thus only trade at scheduled times, as opposed to the continuous trading of exchanges or other ATS. Since traders enter unpriced buy or sell orders, crossing systems do not contribute to price discovery. Execution risk remains at crossing networks since the trade is not necessarily executed. The intuition is that excess demand or supply may result. The advantage of a crossing network is that it minimizes market impact. Trades are typically executed at the midpoint of the bid–ask spread in the primary market. According to SSB, crossing networks cater to institutional investors placing larger sized orders in less liquid securities. Examples of crossing networks for Europe include ITG's POSIT or E-Crossnet. Other crossing systems use an auction procedure (e.g. Arizona Stock Exchange). They are similar to the batch auctions used at traditional exchanges as they match buyers and sellers at the same price in maximizing the matched volume.

A third type of ATS applies Smart Order Routing Technology (SORT). These are systems developed by a variety of market participants that are used to route orders to centralized markets based on trading criteria that seek to provide best execution for the client. This execution can be on a traditional exchange or on an electronic communication network. The trading criteria can be price improvement or execution speed.

ATS are evolving quickly and their future remains quite uncertain. SSB distinguishes several business models for ATS. Some of them move to become a *destination exchange* (e.g. Tradepoint into Virt-X, Archipelago). This implies that the ATS becomes an organized market allowing them to become a destination for listed shares. Another business model is to become a regular broker at several exchanges, i.e. a *destination broker-dealer*. This essentially happens with SORT that should be able to provide execution at several places (e.g. Instinet, a subsidiary of Reuters Company, has become a member at 18 exchanges). It is clear that some new marketplaces offer several of the types of ATS discussed. For instance, ITG is offering an ECN and a crossing network. Moreover, some of the specific aspects of ATS have been already incorporated into the traditional exchanges (NYSE direct+ offers a crossing network).

4 Alternative Trading Systems and liquidity

In this section, we will first discuss US evidence on ATS and liquidity. Next, some evidence for Europe will be presented.

4.1 Empirical evidence for the USA

4.1.1 Importance of ATS

Table 10.4 provides an overview of some main characteristics of the most important ECNs in the USA. The market shares of the ECNs are presented in Figure 10.1.

Most ECNs started operating in the late nineties. Nine of them are currently still active. Jointly, they attract about 29 per cent of total share volume on NASDAQ (second quarter of 2001), a number that has been steadily increasing from 12 per cent in the first quarter of 1998. According to Weston (2001), two causes can be

Table 10.4 ECN characteristics (USA)

ECN	*Archipelago*	*ATTAIN*	*Bloomberg Tradebook*	*BRUT/Strike*	*Instinet*	*Island*	*MarketXT*	*NEXTRADE*	*REDIBOOK*
Starting date	01/97	02/98	12/96	05/98	1969	01/97	01/00	11/98	11/97
Ownership by strategic partners	Yes	Not yet	No	Yes	Reuters	Yes	–	–	Yes
Strategic partnerships	Tradepoint	–	–	–	Yes	–	Yes	Yes	–
Technology	Internal book SORT Plans to form exchange	Internal book	Internal book SORT Agency broker	Internal book SORT	Internal book Agency broker Block trades	Internal book	Mainly after-hours trade SORT	Internal book SORT Exchange application pending	Internal book SORT
Trade volume as percentage of NASDAQ (last quarter 2000)	1.9%	0%	1.3%	1.8%	13%	7.1%	0%	0%	1.7%

Source: Internet.

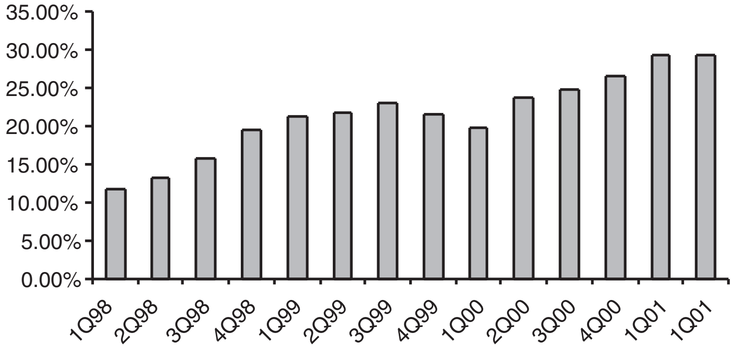


Figure 10.1 NASDAQ share volume of ECNs.
Source: NASDAQ and websites of ECNs.

discerned for this growth pattern. First, the changing SEC regulations are an important determinant. For instance, the so-called 'Order Handling Rules', introduced in 1997, increased competition because public limit orders were since then allowed to compete directly with NASDAQ market makers. Also, market makers posting orders on ECNs were since then obliged to make those orders available for the public as well. This forced dealers to provide greater access to ECNs for public investors. Moreover, ECNs have been more successful in attracting trade from NASDAQ. The intuition is that ECNs offer an agency alternative, eliminating the spread charged by dealers. The NYSE is already an auction market (with a specialist) and enjoys an incumbency advantage due to the liquidity externality. Second, the advances in technology have also played a tremendous role. As argued in the previous section, the trading systems were less advanced compared to many European exchanges. This allows the ECNs to attract a significant part of the market.

An interesting feature of ECNs is that important broker-dealers have become important shareholders of ECNs. Milbourn, Boot and Thakor (1999) offer one explanation for this evolution.[9] Although the future of ECNs is highly uncertain, broker-dealers are willing to accept relatively low returns for the moment. The potential vital role of ECNs in the future defines these ownerships as a strategic option. Moreover, attracting order flow by institutional investors or retail investors, ECNs develop strategic partnerships with online brokers and broker-dealers. This ultimately determines the success of their business plan. All ECNs offer at least an internal limit order book. Most of them also offer SORT; that is they route orders to either other ECNs or dealers. Two of the ECNs have an exchange application pending. The most successful ECN in terms of NASDAQ stocks is Instinet. It attracted about 13 per cent of NASDAQ share volume during the last quarter of 2000. Another important player in terms of NASDAQ volume is Island with 7 per cent. The other ECNs have a market share in NASDAQ volume of less than 2 per cent.

Next to the nine ECNs, there are some crossing networks already operating in the USA and some are announced. Since 1990 the Arizona Stock Exchange organizes single price auctions four times a day. Its volume however is fairly low. ITG's POSIT was launched in 1987 and crosses seven times a day. It is the largest crossing network in the USA. Two very recent crossing systems are Primex Trading and Wofex. Primex Trading exposes certain orders for auction-style competition. Prices may be away from the best quotes in the National Market System. Wofex essentially adds SORT to a crossing network.

4.1.2 Market quality/liquidity

For the United States, there are already some studies describing the behaviour of ECNs and their impact on the market quality/liquidity of traditional exchanges. These include the following: Huang (2000), Hendershott and Mendelson (2000), Simaan, Weaver and Whitcomb (1998), Domowitz (2001), Barclay, Hendershott and McCormick (2001), Weston (2001), Conrad, Johnson and Wahal (2001), Benhamou and Serval (2000), Domowitz, Glen and Madhavan (2001), Domowitz and Steil (1999) and Naes and Odegaard (2001). In most of these studies, the traditional market under consideration is NASDAQ as ECNs have proven to perform best for shares noted on this exchange (see below). In this subsection, we will briefly describe and compare the main results of some of these studies. This will be done focusing on four aspects of market quality/liquidity, namely bid–ask spreads, depth of the market, informational efficiency and price discovery.

BID–ASK SPREADS

Weston (2001) investigates whether the increased market share of ECNs leads to tighter spreads (monthly average quoted, effective and relative spreads for stock *i* in month *t*), i.e. whether ECNs have a significant negative impact on spreads on traditional markets. For this purpose, he performs the following regression using a long time series and large sample of firms:[10]

$$\begin{aligned}\ln(\mathit{Spread})_{it} = {} & \alpha_i + \beta_1(\mathit{ECNdummy})_{it} + \beta_2\ln(\mathit{ECNshare})_{it} + \\ & \beta_3\ln(\mathit{Reforms})_{it} + \beta_4\ln(\mathit{Size})_{it} + \beta_5\ln(\mathit{Turnover})_{it} + \\ & \beta_6\ln(\mathit{Volatility})_{it} + \beta_7(\mathit{numberoftrades})_{it} + \varepsilon_{it}\end{aligned}$$

The variables *ECNdummy*, indicating ECN activity in a given stock-month, and *ECN share* test the effect of ECN activity on spreads. The variable *Reforms* is included to capture possible spread effects of any market reforms (i.e. order handling rules). The independent (control) variables in this model were chosen according to Stoll (1992) and Wahal (1997). They are used to capture well-known determinants of bid–ask spreads, and of execution costs in general. For instance, the selected size variable controls for the fact that orders that are large relative to normal trading volume are likely to have higher execution costs because of adverse selection effects. Log transformations of these variables are used to reduce the skewness.

The coefficients β_1 and β_2 are of interest to us and are consistently negative and statistically and economically significant for all specifications (i.e. for the three kinds of spreads). This implies that ECNs induce competitive pressure on the NASDAQ market. The exact value for the coefficient β_1 in the average quoted spread regression is equal to –0.0041, implying a 4 per cent *ceteris paribus* decrease for this spread measure. For the effective and the relative spread, this decrease amounts to 10 and 7 per cent respectively. The coefficient β_2 indicates that a 1 per cent increase in ECN activity lowers the average quoted spread by 0.714 per cent. For the effective and the relative spread, this decrease amounts 0.917 and 0.07 per cent respectively. Weston argues that these results are particularly strong because the data used actually give an underestimation of the true impact due to the manner in which volumes are reported to the NASDAQ (see above). Note, however, that they are only valid for small trades, not for block trades.

Thus, in addition to regulatory market reforms, the growth of ECNs has helped to significantly lower trading costs. As such, it has mitigated the negative effects of the suspected imperfect competition among NASDAQ dealers (e.g. Huang and Stoll, 1996; Christie and Schultz, 1994 and Weston, 2000).[11]

Domowitz (2001) constructs an American sample by gathering data from institutional investors. For this dataset, total trading costs for executions by institutional investors through ECNs and through traditional brokers and markets are compared.[12] Calculated yearly savings from 1993 through 1996 using automated systems vary from 31 to 65 per cent, relative to trades executed by traditional brokers or dealers.[13] Domowitz even manages to invalidate the conventional wisdom that automated trading venues are cheaper only because 'easier' trades are more often sent to them, as he proves that even for more difficult trades, savings from automated execution are evident.[14]

This empirical evidence is also consistent with Conrad, Johnson and Wahal (2001), although they use a somewhat different approach. They determine what the difference in realized execution costs is between external crossing systems (POSIT or an after-hours cross on Instinet), ECNs (Instinet) and traditional markets (NYSE, Amex or NASDAQ). These three trading systems are engaged in a competition for order flow. In their dataset,[15] the distinction is made between single and multiple mechanism orders, which are, respectively, orders that are completely executed by a single trading system (91 per cent of all orders) and those in which trades are filled by more than one trading system (9 per cent of all orders). Note that there is considerable time series variation, but no trend in the distribution of single mechanism orders. Further, the data show substantial differences in size between orders executed on the three mechanisms. Order fill rates are lowest for crossing systems as they concern a mere function of liquidity on the system (cf. contra-side depth), which is exogenous to the trader. As traders on ECNs and on traditional broker systems can trade anonymously, they endogenously increase the probability of a fill. Evidently, multiple mechanism orders have the largest execution costs, as they are most difficult to fill.

As in Domowitz (2001), total execution costs are measured as the sum of implicit and explicit costs. Obviously, comparing execution costs between different trading systems univariately can be quite misleading as the trading mechanisms may represent varying degrees of aggressiveness on the part of the institution.[16] One needs to take differences in order characteristics between these systems into account. For instance, variation in order difficulty and other characteristics influencing liquidity and thus trading costs. These are controlled for using two methods, i.e. a 'matched-sample' approach[17] and a regression-based approach[18] as in Weston (2001).[19] Both these methods yield quite similar results. Compared to traditional brokers, execution costs on crossing systems are substantially lower. For ECNs, this cost advantage is even more pronounced. Note that these results are quite robust and that the differences can be primarily attributed to distinct implicit costs.

Conrad, Johnson and Wahal (2001) note, however, that an endogeneity problem may arise as the choice of trading mechanism could be endogenous to (*ex post*) realized execution costs. Orders that are more difficult to fill, and thus incur higher *ex post* execution costs, are more likely to be sent to mechanisms guaranteeing a high fill rate. This issue, which leads to inconsistent estimates, is not accounted for in the above-mentioned methods and therefore needs to be addressed by using a two-stage procedure ('endogenous switching regression method') following Madhavan and Cheng (1997). The cost differentials described above seem to persist when applying this model; in fact they are even more pronounced.

DEPTH OF THE MARKET

Besides performing a bid–ask spread comparison, Weston (2001) also investigates whether the increase in ECN market share leads to greater depths. For this purpose, he performs the following regression:

$$\begin{aligned} Depth_{it} = \; & a_0 + a_1 ECNactivity_{it} + a_2 \ln(volume)_{it} + a_3 \ln(price)_{it} \\ & + a_4 \ln(volatility)_{it} + a_5 \ln(MarketConcentration)_{it} \\ & + a_6 TimeDummy_t + \varepsilon_{it} \end{aligned}$$

The presence of an ECN does seem to increase the quoted depth *ceteris paribus* by 11.6 per cent. A 1 per cent increase in ECN activity improves quoted depth by 0.27 per cent all other variables held constant. So ECN activity improves the total quality of the market. These conclusions, however, are disputed by Barclay, Hendershott and McCormick (2001) who study transactions data for June 2000 and conclude that ECN trading lowers quoted depths.

INFORMATIONAL EFFICIENCY

Weston (2001) suggests that ECNs do impose higher adverse selection costs on traditional markets through more anonymous trading.[20] An increase in anonymity through ECN trading may therefore increase information costs, urging intermediaries to charge larger spreads (Amihud and Mendelson, 1986 and Glosten and Harris, 1988). So, although ECNs lower trading costs (cf. above), they reduce the informational efficiency of prices. Note that this conjecture does not hold if the ECN functions as a separate market. In this case the presence of an ECN reduces the amount of information asymmetry in a dealer market by providing an alternative venue for information-based trades. Weston performs a test on the change in anonymity of trading on the NASDAQ due to ECN trading, i.e. estimating the adverse selection component of spread (Huang and Stoll, 1997) and regressing this measure on the level of ECN activity and a group of control variables.[21] An increase in adverse selection costs linked to ECN trading is noticed, confirming the first conjecture stated above. However, these costs are outweighed by benefit of lower overall transaction costs.

PRICE DISCOVERY

Conrad, Johnson and Wahal (2001) describe the link between the efficiency of the primary markets' price discovery mechanism and the success of ECNs. For the USA, it has been extensively proven that transaction costs are significantly lower on the NYSE than on NASDAQ (for example Hasbrouck, 1995 and Huang and Stoll, 1996). An obvious rationale for this difference is the distinction in trading mechanisms that are employed on both markets, i.e. auction markets provide more adequate price discovery than the dealership markets. In their study, they refer to Hendershott and Mendelson (2000), who state two necessary conditions for crossing systems to be successful when coexisting with a dealer market. First, as these systems do not provide active price discovery themselves, they need to rely on a primary market providing an adequate price discovery mechanism. Second, the crossing network initially needs to attract at least a minimum threshold of volume from this primary market so that the pool of liquidity is sufficiently large.[22] Based on these conditions, one could postulate that crossing networks will be more successful in competing for NYSE shares and therefore primarily focus on listed securities. ECNs, on the other hand, engage themselves in active price discovery, and will therefore rather compete with primary markets with higher transaction costs and fragmented order flow.[23] In fact, their success is inversely related to the efficiency of the primary market, i.e. if bid–ask spreads are higher on the primary market, ECNs become a truly competitive alternative.[24] Clearly, external crossing systems and ECNs compete for order flow in different dimensions as certain client effects arise. Empirical evidence seems to support both these conjectures as 90 per cent of all orders executed on external crossing systems are for NYSE securities and 80 per cent of all ECN-executed orders are for NASDAQ securities (sample by Conrad, Johnson and Wahal, 2001).

4.2 Empirical evidence for Europe

It is a much more difficult exercise to gauge the importance of ATS in comparison to the European exchanges. ECNs, like Instinet, are broker-dealers allowing investors to trade on several European exchanges. The two most prominent ECNs that are active on the European market are Tradepoint and Jiway.

1 *Tradepoint*: Tradepoint recently merged with Swiss Exchanges into Virt-X, which is an attempt to create a pan-European blue chip exchange. Using unprecedented technology, their aim is to become competitive by providing the scope for significant reduction in cross-border transaction costs at each stage of the trading, clearing and settlement process. Trading on a sectoral base is encouraged, rather than trading on a national base. Their aim is to 'capture ten per cent of the pan-European blue chip trading within twelve months' (cf. site www.virt-x.com). Actually the market's structure (i.e. a continuous electronic public limit order book with opening, intra-day and closing single-price auctions and full anonymity and facilities to support liquidity providers and off book and block trading requirements) strongly resembles the one offered by auction markets (see below).

2 *Jiway*: Jiway, an initiative of OM Grüppen and Morgan Stanley Dean Witter, combines a limit order book and market makers.[25] Its major focus is giving retail investors greater access to European and American stock markets. Thus it aims at small orders and tries to internationalize the retail market so as to improve liquidity on these markets.

Next to the ECN, there are at least two crossing networks active in European stocks, i.e. ITG-Europe's POSIT and E-Crossnet. It is difficult to obtain estimates of their activity, but, in essence, ATS in Europe are far less important than in the USA.

1 *ITG-Europe's POSIT*: Since 1998, institutional investors and broker-dealers can trade on ITG Europe's POSIT, a crossing system that is active in shares of eight European countries. Anonymity is guaranteed to reduce market impact.

2 *E-Crossnet*: Since 1999, E-Crossnet operates in fourteen European countries and also aims at institutional investors and broker-dealers. Its objectives, structure and dealing mechanisms are roughly comparable to those of POSIT.

The empirical evidence on the interaction between ATS and market quality for Europe is rather scarce. Board and Wells (1999) offer a comparison of SETS (traditional exchange) and Tradepoint (ECN) concerning liquidity and best execution of UK shares. Therefore they compare prices available on those two exchanges, in fact the extent of price improvement opportunities is measured and analysed.[26] Their analysis indicated that while SETS was clearly more active during the period under consideration, Tradepoint managed to offer better prices for between 45 and 90 minutes per trading day, at volumes that were roughly comparable to those offered by SETS. The reason why they still did not manage to attract sufficient trading volume, although being cheaper, is attributed to insufficient depth. Board and Wells propound that 'if the other ECNs that are operating or planning to operate display similar results as Tradepoint, and particularly if they attract significant business, then there will be significant periods of the day in which the SETS price is not the most attractive price'.

Note that most of the empirical evidence for the impact of ECNs on traditional exchanges for the USA concerns NASDAQ (cf. above), which is a dealer market. Therefore, the stated findings and insights do not necessarily apply for Europe where most of the stock markets offer an electronic auction-based mechanism. As this sort of trading system is quite analogous to the one offered by ECNs, they will probably even face difficulties in attracting trade volumes. Clearly, the traditional markets, acting as incumbents, enjoy a major liquidity externality, implying difficulties for ECNs in capturing a market segment of their own. Moreover, it is clear that ECNs have not flourished to such an extent in the European markets so far, because the traditional exchanges have been proactive in addressing the changing needs of investors, i.e. in creating efficient trading facilities themselves.

Intuitively, we expect crossing networks to be relatively more successful than ECNs for Europe. To reach this conclusion, we extend the Conrad, Johnson and Wahal (2001) results on price discovery to the European 'auction market' case. Unfortunately there is no evidence reported so far concerning the magnitude of trading on crossing networks. Initiatives such as E-Crossnet and ITG Europe's POSIT (cf. above), however, demonstrate that these kinds of networks can indeed be erected. Given these arguments, one could expect the traditional markets to create their own passive call market in the future, parallel to their own market.

In the market microstructure literature, previous studies have investigated whether the trading activity of a dually traded stock on one market has an effect on trading activity on the other one (and thus not necessarily on the spread). Pagano and Röell (1991) initiated this research methodology and investigated whether trading of Italian equities on SEAQ International implied trade diversion or trade creation for the Milan Stock Exchange. This methodology has been replicated by, for example, Anderson and Tychon (1993) and Degryse (1996) for the impact of SEAQ International on Belgian equities.

We will now apply this methodology to test whether a variable related to trading activity on an ECN (Virt-X) helps to explain trading volume on the 'local' exchange (Paris Bourse). Trading on ECNs may have displaced activity from the local exchange to the ECN. Alternatively, it may also have generated a stimulus in trading as some institutional investors find the source of or outlet for the shares dealt at the ECN. Our dataset consists of weekly trading data for an eight-month period on the Paris Bourse and on Virt-X. Functioning as dependent variable for our regressions is the volume on the Paris Bourse exchange for ten randomly selected dually traded stocks that are all members of the CAC40 index.[27] Next to the Virt-X volume variable, explanatory variables are lagged values of the dependent variable, current and lagged values of total market volume, current and lagged values of the average return and the volatility of the relevant stock (respectively measured by the monthly average and the standard deviation of daily returns).[28] These are included to control for other possible determinants of trading volume, a choice that is based on Pagano and Röell (1991). A negative and significant coefficient for the Virt-X volume variable is interpreted as a symptom of trade diversion from the Paris Bourse to Virt-X. A positive significant coefficent indicates trade creation. Most of our regressions (nine out of ten), however, generate a negative but insignificant coefficient indicating no effect at all. Note that this is probably due to the small time span of the sample. We expect the effect to increase, within certain boundaries (cf. above), as Virt-X will continue to be operative. Clearly a period of almost three months, in which its introduction to the financial markets occurred, is rather short to state any strong conclusions on its impact. Moreover, our estimated coefficients could also capture some sort of 'summer effect', as the introduction of Virt-X coincides with the summer break. This obviously limits our results.

5 Concluding remarks

The purpose of this paper is to discuss the relationship between alternative trading systems (ATS) and liquidity. Two important trends can be distinguished. First, ATS are currently more successful in the USA than in Europe. Second, within the USA there exists an interesting divergence between the impact of ATS on the NASDAQ dealer market and on the NYSE. ATS are attracting about 30 per cent of market share in the NASDAQ market, whereas their impact on the NYSE is rather small. Trading volume on ATS in Europe is currently still marginal compared to the established marketplaces.

Two forces explaining these differences should be distinguished. The first is that European traditional marketplaces were automated earlier than their American counterparts. International evidence shows that automation reduces transaction costs considerably. ATS are the exponent of automated systems and should therefore be more successful in the USA. Our empirical work shows that automation also has a significant impact on trading costs in Europe, but still less substantial than in an international context. This observation brings us to a second explanation, i.e. the agency nature of trading. European markets are mostly organized as an auction market where traders can submit market and limit orders. ECNs allow investors to trade with each other via a limit order book without the intervention of a dealer. This market microstructure is close to the one of incumbent European exchanges (e.g. Euronext). Therefore ECNs are successful in attracting NASDAQ trading volume and are expected to be less successful in competition with the NYSE or European exchanges. Crossing networks are more successful in realizing trades of NYSE listed securities. This leads us to the projection that crossing networks may be a more successful ATS business model in Europe than ECNs.

Several studies discuss the impact of ATS on the market quality/liquidity of American markets. Bid–ask spreads seem to decrease due to competition of electronic communication networks. Thus competition seems to be more important than fragmentation of markets. The results on market depth are inconclusive. ECNs reduce the informational efficiency of the market. The reason is that ECNs typically allow for anonymous trading, leading to an increase in the adverse selection component of the spread. Crossing networks rely on price discovery at the primary exchange while ECNs actively contribute to the price discovery process.

Currently, trading volume on Alternative Trading Systems in Europe is rather low compared to the established marketplaces. Consistent with this, our empirical work does not reveal significant trade diversion or trade creation effects of Virt-X on the incumbent European exchanges. Evidence from the interaction between Tradepoint and SETS shows that ATS may face a problem of market depth in Europe.

Notes

1 Elkins/McSherry receive trade data on all global trades by institutional traders and compute measures of trading costs. The data consist of average total trading costs – execution commissions, fees and market impact (difference between the price of a stock trade and the average of that stock's high, low, opening and closing prices during the day) – as a percentage of trade value for active managers in a universe of 42 countries.

2 Trading floor development costs, for instance, were calculated to be two to 40 times as expensive as those for electronic market places.

3 On automated markets the electronic order books are open for inspection to all clients allowing for an optimal active liquidity management to control implicit transaction costs.

4 Other reasons include industry and regulatory related matters.

5 Data are taken from the FIBV website. Volatility is measured as the standard deviation of the monthly returns on the countries' stock index. Turnover is proxied by total trading volume divided by total market capitalization.

6 Other specifications show that the magnitude of the coefficient is fairly stable. However, the electronic market dummy not always turns out to be statistically significant. Moreover, the control variables do not always show the expected signs.

7 A notable exception is Jiway, which allows dealers to be dual capacity traders, i.e. also to trade on their own account.

8 The distinction between the different types of ATS is not always clear cut. For example, electronic communication networks often also offer SORT technology.

9 Note that they apply it in a different context, namely the diversification of banking activities.

10 This is a multivariate fixed-effect model that allows for within-firm variation in the parameters to account for unobserved heterogeneity in liquidity for the sample of firms.

11 This is supposed to be due to practices such as payment for order flow and preferenced trading used by traditional dealers to attract order flow through non-price competition. Thus, large spreads are prevented from being enticed away (Weston, 2001).

12 Note that total trading costs also include price impact, determined as a geometric average of realized and effective spreads, and measured relative to short-run industry performance.

13 Average savings amount to 46 per cent.

14 Domowitz defines more difficult trades as having above-median values of trade size and volatility, or having below average market capitalization (firm size), i.e. the controls used above.

15 Note that only to describe ECN activity, only data for Instinet were used as the remaining ECNs only commenced operations after the end of their sample period.

16 Conrad *et al*. (2001) offer the following ranking on aggressiveness: external crosses < ECN executions < broker-dealer operations. These differences result in a natural sorting of order difficulty across the categories.

17 Which controls for trade direction, order instruction, order size, exchange listing and market capitalization without imposing any functional form restrictions.

18 Control variables: order size, inverse of stock price, logarithm of market capitalization, exchange listing, return volatility, cumulative size-decile adjusted return, institution-specific indicator variables, indicator variables for external crosses and ECN-executed orders.

19 Note that another possibility for comparing execution costs is focusing on multiple mechanism orders, as order characteristics by definition are held constant across the trades. Also the investor chooses how to break up the order, and where and in what sequence to place the order.

20 Intermediaries face uncertainty on the type of trader they deal with, i.e. informed or uninformed ones.

21 These control variables include market capitalization, share turnover, return volatility and market concentration, and are also suspected to affect information costs.

22 Referring to the Hendershot and Mendelson paper, Conrad *et al*. quote that ‘Volume on crossing systems that provide no price discovery function has a natural upper bound since the system cannot exist independent of the primary price-setting mechanism, whether it be an auction or dealer market. To the extent that other systems (such as ECNs) provide a price discovery mechanism, they can exist and grow independently.’

23 ECNs do make a significant contribution to price discovery and therefore do not necessarily engage free-riding of price discovery by traditional dealers on NASDAQ (Huang, 2000).

24 Note that a major determinant of the higher bid–ask spread on NASDAQ is the difference in anonymity, i.e. the NASDAQ market structure is more anonymous than the NYSE (Garfinkel and Nimalendran, 1998 and Heidle and Huang, 2000) leading to higher adverse selection costs and thus to higher spreads.

25 Note that meanwhile (as of September 2001), MSDW sold their stake integrally to OM Grüppen.

26 The comparison is executed towards some specific factors, e.g. the availability of best prices on the two markets, the spread on each market, available depth at best prices, etc.

27 Namely for Alcatel, Aventis, Axa, Carrefour, Eurotunnel, France Telecom, Orange, Renault, Usinor and Vivendi.

28 Note that replacing the Virt-X volume variable by a dummy variable as in Pagano and Röell (1991) does not change our results.

References

Amihud, Y. and Mendelson, H. (1986) ‘Asset pricing and the bid–ask spread’, *Journal of Financial Economics* 17: 223–49.

Anderson, R. and Tychon, P. (1993) ‘Competition among European financial markets: the case of cross-listed Belgian equities’, IRES Discussion Paper 9314.

Barclay, M., Hendershott, T. and McCormick, T. (2001) ‘Electronic communications networks and market quality’, Working Paper, University of Rochester.

Benhamou, E. and Serval, T. (2000) ‘On the competition between ECNs, stock markets and market makers’, Financial Markets Group Discussion Paper.

Board, J. and Wells, S. (1999) ‘Liquidity and best execution in the UK: a comparison of SETS and Tradepoint’, Financial Markets Group Discussion Paper.

Conrad, J., Johnson, K. and Wahal, S. (2001) ‘Alternative trading systems’, Working Paper.

Christie, W. and Schultz, P. (1994) ‘Why do NASDAQ market makers avoid odd-eighth quotes?’, *Journal of Finance* 49: 1813–40.

Degryse, H. (1996) ‘Competition between auction and dealership markets: Brussels *versus* London’, *Revue de la Banque/Bank-en Financiewezen*, 378–86.

Demsetz, H. (1968) ‘The cost of transacting’, *Quarterly Journal of Economics* 82: 33–53.

Domowitz, I. (2001) ‘Liquidity, transaction costs, and reintermediation in electronic markets’, Working Paper.

Domowitz, I., Glen, J. and Madhavan, A. (2001) ‘Global equity trading costs’, Working Paper.

Domowitz, I. and Steil, B. (1999) ‘Automation, trading costs, and the structure of the securities trading industry’, Brookings-Wharton Papers on Financial Services.

Garfinkel, J. and Nimalendran, M. (1998) ‘Market structure and trader anonimity: an analysis of insider trading’, Working Paper.

Glosten, L. and Harris, L. (1988) ‘Estimating the components of the bid–ask spread’, *Journal of Financial Economics* 21: 123–42.

Glosten, L. and Milgrom, P. (1985) 'Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders', *Journal of Financial Economics* 14: 71–100.

Hasbrouck, J. (1995) 'One security, many markets: determining the contributions to price discovery', *Journal of Finance* 50: 1175–99.

Heidle, M. and Huang, D. (2000) 'Information-based trading in dealer and auction markets: an analysis of exchange listings', Working Paper.

Hendershott, T. and Mendelson, H. (2000) 'Crossing networks and dealer markets: competition and performance', *Journal of Finance* 55: 2071–116.

Ho, T. and Stoll, H. (1981) 'Optimal dealer pricing under transactions and return uncertainty', *Journal of Financial Economics* 9: 47–73.

Huang, R. (2000) 'Price discovery by ECNs and NASDAQ market makers', Working Paper, University of Notre Dame.

Huang, R. and Stoll, H. (1996) 'Dealer vs. auction markets: a paired comparison of execution costs on NASDAQ and the NYSE', *Journal of Financial Economics* 41: 313–57.

Huang, R. and Stoll, H. (1997) 'The components of the bid–ask spread: A general approach', *Review of Financial Studies* 10: 995–1034.

Madhavan, A. (2000) 'Market microstructure: a survey', *Journal of Financial Markets* 3: 205–58.

Madhavan, A. and Cheng, M. (1997) 'In search of liquidity: block trades in the upstairs and downstairs markets', *Review of Financial Studies* 10: 175–203.

Milbourn, T., Boot, A. and Thakor, A. (1999) 'Megamergers and expanded scope: theories of bank size and activity diversity', *Journal of Banking and Finance* 23: 195–214.

Naes, R. and Odegaard, B.A. (2001) 'Equity trading by institutional investors: to cross or not to cross?', Working Paper, Norwegian School of Management.

Pagano, M. and Röell, A. (1991), 'Dually-traded Italian equities: London vs. Milan', CEPR Discussion Paper 564.

Salomon Smith Barney (2001) 'Trading up: The new world of electronic trading', Research Report.

SEC (2000) *Securities and Exchange Commission Report*.

Simaan, Y., Weaver, D. and Whitcomb, D. (1998) 'The quotation behavior of ECNs and NASDAQ market makers', Working Paper, Baruch College.

Stoll, H. (1992) 'Principles of trading market structure', *Journal of Financial Services Research* 6: 75–106.

Wahal, S. (1997) 'Entry, exit, market makers, and the bid–ask spread', *Review of Financial Studies* 10: 871–901.

Weston, J. (2000) 'Electronic communication networks and liquidity on the Nasdaq', working paper.

Weston, J. (2001) 'Competition on the NASDAQ and the growth of electronic communication networks', Working Paper.

Acknowledgement

We appreciate helpful discussions with Jean-Paul Abraham, Jos Schmitt and participants at the 23rd SUERF Colloquium. We acknowledge financial assistance from FWO-Vlaanderen under contracts G.0302.00 and G.0333.01.

11 Where traders go when stock exchanges go virtual – concentration, dissemination or persistence?*

Vivien Lo and Michael H. Grote

Abstract

While the impact of technology on different market participants such as stock exchanges and banks has gained a prominent position on the research agenda, the geographical consequences of the new communication technologies are mostly overlooked. However, for the existence of financial centres these questions are vital: Where will traders locate when stock exchanges finally become pan-European and accessible from any place? Will traders disseminate to a random distribution, will they concentrate in any place, and if so, where? What are the underlying rationales and mechanisms of these potential movements?

With the possibility of remote access, the need for the on-site presence of traders seems to disappear. Several urbanization and localization effects in the financial sector vanish, while others increase in importance. The empirical evidence of the agglomeration of traders, i.e. predominant financial centres in almost every country, shows that there are strong tendencies for traders to cluster. However, it is not clear whether these tendencies are strong enough to overcome barriers of language, culture and regulatory usances.

This paper argues that with the virtualization of stock exchanges, informational spillovers with other traders and the headquarters of traded firms become the main determinants of the location of stock traders. When the relative strength of these two forces varies over time the final location of traders may remain indeterminate.

1 Introduction: changing financial landscapes

Like the financial sector in general, stock markets tend to concentrate in specific locations: almost every nation has only one leading stock exchange, with the others left far behind. Frankfurt's stock exchange is synonymous for the German stock market, Paris for the French, London for the British, Tokyo for the Japanese, and so on. In all of the examples, other national stock exchanges have existed or still exist but are of minor importance. At first glance, the USA seems to be an exception, with NASDAQ and Wall Street as two leading stock exchanges. However, a closer look reveals that each has specialized on a specific market seg-

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

ment (technology-related shares, blue chips) – and that they are the leading stock exchanges in their respective segments by a large margin.

In former times, physical presence at the stock exchange was absolutely necessary to execute trades on the floor. As traders are closely associated with banks, it is not surprising that the rankings of stock exchanges within countries closely reflect the rankings of financial centres. Most national financial business takes place in the previously mentioned cities, and the majority of national financial actors are located there, again by large margins. Financial centres and stock exchanges emerged in times when national financial systems were rather isolated from other countries. But borders are vanishing, most pronouncedly in Europe.

The introduction of the euro, deregulation, and especially the development of new information and communication technologies have greatly impacted the European financial system (Schamp, 1999), making it almost a 'level playing field'. The organization of the trading process is shifting away from floor-based trading to screen-based systems. With 'transportation costs' approximating zero, it is now possible for spatially separated traders to tap directly into the same market. Most incumbent stock exchanges and virtually all new players like the ECNs (electronic communication networks, e.g. Jiway, EASDAQ and Tradepoint) are introducing technologies that allow remote access to all participants. This leads to the question: What will happen in Europe? Will traders become completely footloose, initiating the age of the 'end of geography'? Or will they concentrate in one dominating stock exchange? What are, then, the forces drawing traders to one place or the other?

The paper is organized as follows. After a short look at the mechanisms of stock exchange competition in Section 2, factors that lead to the clustering of traders are analysed in Section 3. Section 4 deals with the impact of virtualization on the various agglomeration effects, and Section 5 examines the changes in the relative importance of the different forces that influence the location of traders; Section 6 concludes.

2 Stock exchanges and network externalities

To begin with, a brief look at the underlying mechanisms of the current competition between European stock exchanges is in order. Surprisingly, it is not easy to define a stock exchange (Lee, 1998). For our purposes it is sufficient to take the basic functions of stock exchanges into account, namely the matching of buy and sell orders of market participants. Both buyers and sellers prefer markets with a high number of participants and therefore high liquidity, because it reduces the risk of arbitrary prices (Economides, 1995). This leads to a remarkable feature of stock exchanges, namely network externalities.[1] For each actor the utility of using the market increases with every new participant who joins (Benhamou and Serval, 2000; Economides, 1993; Pagano, 1989). Faced with the question of which stock exchange to join, one with a small number of participants or one with a large number of participants, a new actor will – *ceteris paribus* – choose the already larger exchange (Heinemann and Dönges, 2001; see Katz and Shapiro, 1985). This further increases the attractiveness of the already larger stock exchange. Clearly, stock exchanges are subject to increasing returns to scale (Economides and Siow, 1988).[2] Assume that there is a stock exchange in competition with others: the larger the market share of that stock exchange, the higher the probability that a new user will choose it (see Arthur, 1994). This situation[3] is illustrated in Figure 11.1.

This graph is a frequency distribution converted to cumulative form. It demonstrates the relationship between the market share of the stock exchange and the probability of a new participant choosing this stock exchange. To give an example: at 30 per cent market share, the probability is lower than 0.3 (i.e. below the 45° line) indicating an unstable situation. The system is in equilibrium only when the market share of the stock exchange equals the probability of choosing it, i.e. on the 45° line. The ‘s’-shaped graph therefore highlights the importance of a ‘critical mass’ (Economides, 1993): after reaching more than 50 per cent market share, a stock exchange should draw the whole market. Thus, there are only two stable equilibria: either all participants use this stock exchange, giving it a market share of 100 per cent (point B) – or none use it, implying a market share of zero per cent (point A).

The range around 50 per cent market share is of special interest. The equilibrium in point (X) is unstable. Both stock exchanges have a 50 per cent market share, and the probability that a new participant chooses one of the stock exchanges is 0.5 as well. Small deviations from that point – which could be purely random – rapidly lead to one of the other two stable equilibria (see Schelling, 1978). Network externalities can lead to ‘lock-in’ effects. The stock exchange that initially gained a superior market share will attract the whole market even if it has

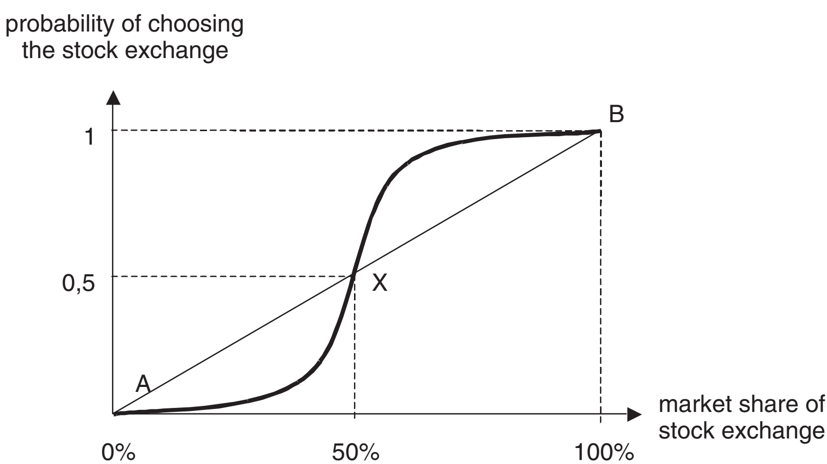


Figure 11.1 Proportion/probability – mapping.
Source: Arthur 1994, p.18.

an inferior technology (which could result in a welfare loss). In theory at least, 'no network of size smaller than this positive size, called critical mass, is ever observed' (Economides, 1993, 92), i.e. the norm should be only one prevailing stock exchange per market. Whether this result is obtained through an outright monopoly or via implicit mergers between stock exchanges is subject to discussion (see Wahrenburg, 2001; Di Noia, 2001) but will not influence the results presented here. Yet, for a variety of reasons (some of them purely political), we see small stock exchanges with very small market shares in almost every country. However, according to these mechanisms it is very likely that in an integrated Europe, liquidity will concentrate in one or very few stock exchanges only.

Since it is expectations about future market shares that drive potential participants to choose a stock exchange, there is some possibility of a 'self-fulfilling prophecy'. Whenever the majority of actors believes that a stock exchange will have a superior market share in the future, participants will choose that stock exchange. A good example for the 'flipping' of market share is the Bund-future, the most traded derivative in the world (see Laulajainen, 2001). Because it is based on German federal government bonds, it could be expected that it would be traded in its 'home country', Germany. However, as German regulation did not allow trading of derivatives until 1990 – it was in fact considered gambling and therefore illegal – the market emerged in the 1980s at LIFFE (London International Financial Futures and Options Exchange) and the relevant expertise concentrated in London. The efforts of the German Derivatives Exchange – predecessor of Eurex – to transfer Bund-future trading volume to Frankfurt met with little success, despite comparatively low trading costs. Due to the above-described 'lock-in effects' – i.e. liquidity – a large extent of the trade still took place at London's LIFFE.

However, the installation of terminals with remote access facility in London provided an easy switch option for London traders. After a combined shift of the derivative trading activities of the large German banks – induced by 'political' lobbying – Eurex gained a critical mass of market share and London traders also began using the Frankfurt-based system. Eurex's market share today amounts to 99.9 per cent (Grote, Lo and Harrschar-Ehrnborg, 2002). With only one stock exchange left executing virtually the whole trade, the so-called 'battle for the Bund-future' is a good example of the mechanisms demonstrated in the model above.

However, this might have been a Pyrrhic victory for the financial centre in Frankfurt: although Frankfurt-based Eurex is now the dominant marketplace for Bund-futures, most of the traders are using the system from foreign locations, mainly London (Seifert, 2000). Frankfurt could thus be called a 'leader in technology, not in people' (Grote, Lo and Harrschar-Ehrnborg, 2002). Since the regional advantages of a central financial marketplace are mainly related to the income generated in that financial centre, the location of traders is crucial for the respective cities. In a way, the new remote access technology transformed liquidity from a spatial agglomeration effect into a virtual network effect. In principle at least, the most liquid market can now be tapped from anywhere through telecommunication networks – at first look, a spatial dissemination of traders seems highly likely.

3 Agglomeration effects in trading: localization and urbanization

Liquidity is not the only reason why traders tend to concentrate spatially, close to a stock exchange in a financial centre. Pull factors that draw traders together in certain locations are generally summarized under the heading of agglomeration effects. Agglomeration advantages are benefits that a single firm derives simply because it is located in spatial proximity to other firms. As these external economies of scale (also termed positive spatial externalities or spillovers) increase with every new participant, they tend to be dynamic (Thrift, 1994). In contrast to other industries, congestion is of less importance to the financial sector, so bigger is literally better. In the following section, the strength of agglomeration economies as pull factors are examined for the case of financial firms or departments specialized in securities trading. The main argument of this paper is that the diffusing use of new information and communication technologies is transforming these centripetal forces. In our analysis, we show that while some of the reasons for the spatial concentration of traders persist even in the age of telecommunication, other effects can be 'virtualized'. Virtualization is understood here as the substitution of the need for spatial proximity by virtual proximity (Bathelt, 2000) via electronic systems.

Two forms of agglomeration economies are generally distinguished: localization economies (Marshall–Arrow–Romer effects) are advantages emerging out of the proximity to firms of the same sector while urbanization economies (Jacobs effects) occur in close distance to firms of other sectors (Fujita and Thisse, 1996). The latter are the general external effects of locating in a city with a diversity of sectors and infrastructure. The so-called 'global cities' of New York, London and Tokyo (Sassen, 1991) feature both, localization and urbanization effects. These economies of scale and scope across firms in the same locale can generate significant comparative advantage to financial firms (Budd, 1998). In the following section, these advantages are analysed in detail.

3.1 Localization economies

Liquidity and price information

As argued above, market liquidity is one of the most important centripetal forces for traders in financial securities. Risk-averse investors prefer to trade in a liquid market, because the risk of price changes caused by individual traders is lower, as is the risk of shocks (Gehrig, 1998). Both liquidity and efficiency grow with the number of participants (Laulajainen, 1998). Until recently, physical presence was required at each stock exchange in order to participate in the local market. It was also the only way to learn current stock prices and to understand the complex and partially unwritten rules of local stock exchange dealings (Grote, Lo and Harrschar-Ehrnborg 2002).

Lower costs of infrastructure use

Closely related to this are external economies of scale through sharing infrastructure like settlement and payment systems. The greater number of participants, the lower each individual's share of these fixed costs of running financial markets becomes (Thrift, 1994). In the past, participation in these systems required a presence on-site for paper exchange. A study among foreign banks in Frankfurt shows that the German money settlement system was one of the major reasons for establishing offices in the city (Grote, 1998).

Informational spillovers

A concentration of financial actors entails greater information turnover. As the number of possible contacts rises with an increase in the number of local actors, it can also be assumed that this local concentration leads to greater connectivity. It is therefore more likely that information and knowledge are circulated, and that their diffusion rate is higher (Porteous, 1999). This in turn leads to the creation of new knowledge and product and process innovations (Thrift, 1994). Because it can be assumed that firms have different levels and types of knowledge, the benefit of communication grows with the number of participating firms. Communication is regarded as distance-sensitive, which implies that the benefits of communication are larger when firms locate close to each other. Traders are dependent on a constant and rapid input of reliable information. Close contact prevents misunderstandings and allows for mistakes to be remedied quickly (Davis, 1990).

Labour market

As the financial sector is still very much a 'people's business' with a high demand for specialized expertise, the labour market argument is generally considered to be a very strong one (Porteous, 1999). A large local pool of specialized labour is not only attractive for firms but also for employees. Both sides can expect to make better matches, because of the heterogeneity of qualification profiles and requirements (Kim, 1991). Furthermore, the agglomeration of a large number of firms and potential employees reduces cyclical variations at the firm level.

3.2 Urbanization economies

Access to intermediaries[4]

As argued earlier, agglomeration economies not only emerge from proximity to actors from the same sector but also those of other sectors. Veltz (1996), for example, emphasizes the general advantages of locating in large metropolitan areas. The greater number and variety of firms can act as insurance against fluctuation and shocks. Veltz argues that this insurance increases in importance as uncertainty in the environment of firms grows. Because financial firms are users of specialized producer services, the concentration of firms from the areas of law, accountancy, consultancy and computers is especially attractive (Lo and Schamp, 2001). Location near the source of such inputs can ensure better service and lower prices (Thrift, 1994). Close proximity is important, because the timeliness of these services can be vital for exploiting profit opportunities (Porteous, 1999). Although intermediary services are often less visible than financial firms, they are indispensable for the proper functioning of a financial centre (Laulajainen, 1998). Thus, the competitiveness of a financial agglomeration might be seen as dependent upon access to knowledge of specialized service suppliers (Lo, 2001). The argument is therefore very close to informational spillovers.

Proximity to other industries

Close proximity and relationships to actors from diverse sectors provide traders with background information and industry insights to which they might not otherwise have access. Informal networks can be used for verifying rumours, so that proximity yields if not more, then more trustworthy, information and thus can lead to greater trading profits. Empirical evidence shows that local proximity of traders to corporate headquarters is positively correlated with intra-day trading profits. This might be taken as an indicator for the existence of local information advantages (Hau, 1999, 2001a).

4 The effects of virtualization

The financial sector is generally considered the most globalized of all economic sectors, 'since the very fungibility and convertibility of money enable it to transcend space more readily than any commodity' (Martin, 1999, 6). Money is easily digitalized and is therefore highly mobile via information and communication technology. For this reason, it is generally believed that the spread of telecommunications will have an enormous impact on the financial sector and financial centres. This paper argues that telecommunications are indeed changing the locational choices of financial actors. However, it is unlikely that traders will turn as footloose and virtual as some authors suggest (see, for example, O'Brien, 1992). We argue that although some agglomeration forces have lost weight, it does not follow that concentration tendency as such is weakened. Instead, pull factors that are less affected by information technology are now emerging as more prominent. They gain in significance, because two factors that in the past acted as strong centripetal forces for traders have changed dramatically with the spread of telecommunication: liquidity and cost of infrastructure.

With the introduction of computer-based trading and settlement and payment systems with remote access, both benefits, liquidity and low cost of infrastructure, do not require an on-site presence any more. The market can be accessed from any location anywhere in the world. Spatial proximity is substituted by virtual proximity on the net. Although the benefits of large numbers still occur, they can be reaped without being tied to one specific location. In these 'virtual agglomerations' the 'agglomeration of traders takes place within electronic communication networks' (Gehrig, 1998, 13). Agglomeration economies thus turn into non-locational network externalities. While these two agglomeration effects have apparently disappeared from a spatial point of view, the impact of telecommunication on information spillovers is not so straightforward. The growth of information exchange via computer networks has been stunning in the last decades and much has been written on the ensuing information age (see, for example, Castells, 1997). Information that used to be restricted and difficult to obtain has now become ubiquitous. Stock prices, for example, are now available on the Internet in real-time. It seems paradoxical that at the same time, access to information is frequently mentioned as one of the most important locational factors.

In the last few years, it has become increasingly evident that there is a crucial difference between information that is readily transferable via telecommunication, and information that is not. These different types of information can be termed straightforward and complex (Gaspar and Glaeser, 1998), standardized and unstandardized (Porteous, 1999) or codified knowledge and tacit knowledge (Cowan, David and Foray, 2000). Contrary to straightforward information, complex information requires face-to-face contact. This is especially the case if the interpretation of the information is highly contingent on different variables or if the information can be easily manipulated. In these situations, a unidirectional information transfer is not enough in order for the receiver to be able to use the information (Lo, 2000). Intervention, clarification and redundancy are necessary for an actor to learn how to process the information and 'to prevent misinterpretations or guarantee a certain degree of confidentiality' (Gehrig, 1998, 33). The transfer of complex information requires not only an interactive but also a 'thick' mode of communication. Body language, voice intonation and eye contact can reveal depths of information of which the written word is not capable (Boden and Molotch, 1994). Costly errors can be prevented by iterations of communication, in order to check the correct interpretation of a message.

Porteous (1999) tackles the same problem when he speaks of the quality of unstandardized information which declines sharply over distance between generator and user. The question of virtual or spatial agglomeration is therefore dependent on the knowledge-intensity of a financial product. Clark and O'Connor follow a similar argument when they introduce three different types of products: transparent, translucent and opaque, depending on whether the information needed is ubiquitous or transaction specific (Clark and O'Connor, 1997). Gehrig (1998) argues that brokers on the same floor exchange information about order books quite frequently. There is a temporary informational advantage that agents on-site can use in contrast to distant brokers who only receive the information when revealed on the price quotes of their computer screens. Interestingly enough, O'Brien (1992, 50) states very similar observations about stock research in his (in)famous book *The End of Geography*: 'There is a range of information that players will not put down in black and white, or onto the screen. Often that information will be obtained only by being close to the market, often physically close to those who know.' The validation of rumours about the state of a firm apparently becomes increasingly difficult with greater distance from the source. Although it is possible to contact local traders for verification, local firms again possess a temporary informational advantage (Agnes, 2000). Not only that, but local traders are also able to use their local connections for getting a 'better and quicker sense of the accuracy and value of new information' (Porteous, 1999, 104).

5 Where traders go …

Although the visions of telecommunication leading to the 'death of distance' or the 'end of geography' seem grossly exaggerated, there is little doubt that the importance of agglomeration economies has changed with the spread of new telecommunication technologies. While virtualization may reduce former locational benefits, it does not follow that location as such is becoming irrelevant. It just might be a different location than before. The main thesis of this paper is that through virtualization, agglomeration effects that used to be less relevant, now gain in significance relative to others that had been key in the past. Before virtualization, liquidity and joint usage of infrastructure used to be the most important attraction factors. While these have clearly lost significance, it does not follow that all agglomeration effects have become devaluated. Now that location near a stock exchange is not a pull factor anymore, other location factors appear from the background, especially informational spillovers (Figure 11.2).

In principle, Electronic Trading Systems allow traders to locate anywhere. However, it is to be expected that so-called 'soft factors' are keeping them from becoming footloose. These soft factors are mainly informational spillovers between different actors. They refer to the need for spatial proximity to:

1 related departments of own bank (i.e. research and portfolio management)
2 other traders
3 headquarters of traded firms
4 headquarters of own bank.

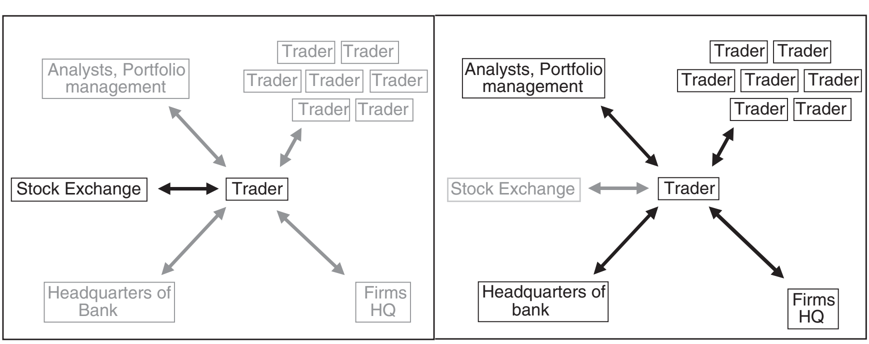


Figure 11.2 Changes in the relative importance of pull factors.

While the last factor refers to the necessity of control, the first three are all related to access to complex, verified and timely information. Proximity to analysts (research) and the sales force is essential, because they are the traders' main source of information. This information is not always clear cut. Often it is more a feeling of how the market is going to move, based on the analyst's experience. In these situations, face-to-face contact is decisive, because it can relay the ambiguous nature of the analyst's assessment. Furthermore, closeness to other traders and company insiders is important in order to establish personal networks. Informal meetings can lead to valuable contacts which pay off in the form of timely information (see Hau, 2001a). All of these contacts are easier to establish and maintain in proximity. Not only are telephone and e-mail more formal than face-to-face, they are also routinely recorded and therefore unsuitable for certain types of information. This does not imply that trading from distant locations is not possible, as, for example, the relatively high share of foreign traders on the Frankfurt Xetra system of 40 per cent shows (Finanzplatz, 2001). However, distance is often connected to a different trading strategy. For a short-term strategy, where the trader has to react rapidly to changes in the market mood, proximity to the aforementioned actors is vital. Traders with a long-term perspective, however, do not necessarily require a listening post in the market.

These findings are supported by Hau (1999, 2001a, 2001b) who studied the effect of distance on proprietary trading profits, and by other literature on the 'home bias' effect of portfolio investment (see, e.g., Portes, Rey and Oh, 2001; Gehrig, 1993) and cross-listing decisions (Pagano *et al*., 2001). Hau used data of traders located in eight European countries with access to Xetra, the electronic trading platform of the German Stock Exchange. The study demonstrates that traders in non-German speaking financial centres show a significant underperformance in large blue chip stocks. As information on blue chips is widely available, it is often assumed that the influence of location is negligible. The study results are therefore highly significant. The results also support the thesis on different trading strategies according to location. In high frequency trading, traders located near corporate headquarters outperform more distant traders in the respective stock. However, there is no discernible difference according to trading location in medium and low frequency trading (Hau, 2001a).

6 Virtual exchanges, virtual agglomeration, what next?

Although space–time compression is discernible in electronic trading, it appears that time and space still make a difference. Traders with a long-term strategy can use virtualization to tap into different markets, while agents with a short-term strategy are still drawn by agglomeration effects. While liquidity and joint use of infrastructure have lost in significance, informational spillovers are moving into the foreground and are increasingly important. This includes informational spillovers within the financial sector (localization effect) as well as between traders – and analysts – and traded firms (urbanization effect). Although the local labour market argument and access to intermediaries have not diminished in importance through virtualization, they do not appear to have increased either. One has to be careful not to underestimate their relevance, although informational spillovers appear to be more prominent.

Yet, the direction of the pull still remains unclear. In the last few years, whole departments of traders have moved from Frankfurt to London and back again (e.g. FAZ, 1999; FAZ, 2001). Not only the trading departments but also the analysts and sometimes even the corresponding part of the board of directors have been shifted between financial centres. Virtualization has set traders free of the pull of the stock exchange, but it does not lessen the importance of the location question. On the one hand, there is the pull of localization to the largest agglomeration of traders, i.e. London in Europe. On the other hand, there is the pull of urbanization to local information on traded companies. With these antipodal forces, it is difficult to project where traders will go in the next years, now that stock exchanges have truly become virtual. If the relative strength of these two forces varies over time, there might not even be a final location. Despite the big regulatory and technological changes that are taking place in Europe, a surprising result is that there will probably not be too much of a change in the location of traders in the foreseeable future.

Notes

1 For a discussion of the distinction between 'network effects' and 'network externalities' see Liebowitz and Margolis (1998). In the following, the above-mentioned definition of network externalities is used as in most literature (see for example Economides, 1993, 1996).

2 There are also other features of stock exchanges, e.g. listing of firms, research, transactions, that are subject to increasing economies of scale (see Di Noia, 2001). For the impact of competition among exchanges on trading standards see Santos and Scheinkman (2001).

3 In order to clarify our point here, we have abstracted the model from the treatment of participants that have already chosen one stock exchange, from sunk costs and from modelling expectations. None of these abstractions change the underlying mechanisms of the competition between stock exchanges, but play a role in possible outcomes and estimated welfare effects (see Liebowitz and Margolis, 1999).

4 The advantage of a local pool of intermediaries are categorized here as urbanization instead of localization effects. Intermediaries for the financial sector, like accountants, lawyers, management consultants, etc., are typically working not exclusively for clients of one sector but for a variety of sectors. Therefore, they play an important role in the knowledge diffusion between different companies and sectors (Lo, 2001).

References

Agnes, P. (2000) 'The "end of geography" in financial services? Local embeddedness and territorialization in the interest rate swaps industry', *Economic Geography* 76(4): 347–66.

Arthur, B.W. (1994) *Increasing Returns and Path Dependence in the Economy*. Ann Arbor: University of Michigan Press.

Bathelt, H. (2000) 'Räumliche Produktions – und Marktbeziehungen zwischen Globalisierung und Regionalisierung – Konzeptioneller Überblick und ausgewählte Beispiele', *Berichte zur deutschen Landeskunde* 74(2): 22–41.

Benhamou, E. and Serval, T. (2000) 'On the competition between ECNs, stock markets and market makers', Financial Markets Group Discussion Paper 345, London School of Economics.

Boden, D. and Molotch, H.L. (1994) 'The compulsion of proximity', in R. Friedland and D. Boden (eds), *Now Here: Space, Time and Modernity*. Berkeley, LA, London: University of California Press, 257–86.

Budd, L. (1998) 'Global cities and finance: a troubled relationship', in S. Gravesteijn, S. van Griensven and M. de Smidt (eds) *Timing Global Cities*, Nederlandse Geografische Studies 241, Utrecht: Koninklijk Nederlands Aardrijkskundig Genootschap, 67–83.

Castells, M. (1997) *The Power of Identity: The Information Age – Economy, Society and Culture*. London: Blackwell.

Clark, G. and O'Connor, K. (1997) 'The informational content of financial products and the spatial structure of the global finance industry', in K.R. Cox (ed.), *Spaces of Globalization*. London, New York: Guilford Press, 89–114.

Cowan, R., David, P.A. and Foray, D. (2000) 'The explicit economics of knowledge codification and tacitness', *Industrial and Corporate Change* 9(2): 211–53.

Davis, E.P. (1990) 'International financial centres – an industrial analysis'. London: Bank of London Working Paper Series.

Di Noia, C. (2001) 'Competition and integration among stock exchanges in Europe: Network effects, implicit mergers and remote access', *European Financial Management* 7(1): 39–72.

Economides, N. (1996) 'The economics of networks', *International Journal of Industrial Organization* 16(4): 673–99.

Economides, N. (1995) 'How to enhance market liquidity', in R. Schwartz (ed.), *Global Equity Markets*. New York: Irwin.

Economides, N. (1993) 'Network economics with application to finance', *Financial Markets, Institutions and Instruments* 2(5) (December): 89–97.

Economides, N. and Siow, A. (1988) 'The division of markets is limited by the extent of liquidity (spatial competition with externalities)', *American Economic Review* 78(1): 108–21.

FAZ – Frankfurter Allgemeine Zeitung (1999) 'Auslandsbanken verlagern wichtige Geschäfte nach London', FAZ 176(2), August 1999.

FAZ – Frankfurter Allgemeine Zeitung (2001) 'Vor allem die Investmenthäuser bauen kräftig aus', FAZ 34(9), February 2001.

Finanzplatz e.V. (2001) *Gateway to Euroland*. Frankfurt/M: Finanzplatz e.V.

Fujita, M. and Thisse, J.-F. (1996) 'Economics of agglomeration', *Journal of the Japanese and International Economies* 10(4), 339–78.

Gaspar, J. and Glaeser, E. (1998) 'Information technology and the future of cities', *Journal of Urban Economics* 43: 136–56.

Gehrig, T. (1998) 'Cities and the geography of financial centres', CEPR Discussion Paper 1894. London.

Gehrig, T. (1993) ‘An information based explanation of the domestic bias in international equity investment’, *Scandinavian Journal of Economics* 95(1): 97–109.

Grote, M. (1998) *Foreign Banks in Frankfurt*, Mimeo. Institute for Social and Economic Geography, Frankfurt University.

Grote, M., Lo, V. and Harrschar-Ehrnborg, S. (2002) ‘Frankfurt’s changing role in the European financial centre system – a value chain approach’, *Tijdschrift voor Economische en Sociale Geografie* 93(4): 412–23.

Hau, H. (2001a) ‘Location matters: An examination of trading profits’, *Journal of Finance* 56(5): 1959–83.

Hau, H. (2001b) ‘Geographic patterns of trading profitability in Xetra’, *European Economic Review* 45: 757–69.

Hau, H. (1999) ‘Who makes profits in Xetra trading? Evidence on trading performance in the German stock market’, Working Paper, Centre for Financial Studies, Frankfurt/M.

Heinemann, F. and Dönges, J. (2001) ‘Competition for order flow as a coordination game’, Working Paper 64: Finance & Accounting, University of Frankfurt/M.

Katz, M. L. and Shapiro, C. (1985) ‘Network externalities, competition, and compatibility’, *American Economic Review* 75(3): 424–40.

Kim, S. (1991) ‘Heterogeneity of labour markets and city size in an open spatial economy’, *Regional Science and Urban Economics* 21: 109–26.

Laulajainen, R. (2001) ‘End of geography at exchanges?’, *Zeitschrift für Wirtschaftsgeographie* 45(1): 1–14.

Laulajainen, R. (1998) *Financial Geography: A Banker’s View*. Gothenburg: Gothenburg University.

Lee, R. (1998) *What is an Exchange?* Oxford: Oxford University Press.

Liebowitz, S.J. and Margolis, S.E. (1998) ‘Network externality’, in *The New Palgrave Dictionary of Economics and the Law*. London: Macmillan.

Liebowitz, S. J. and Margolis, S. E. (1999) *Winners, Losers & Microsoft – Competition and Antitrust in High Technology*. Oakland: Independent Institute.

Lo, V. (2001) ‘Wissensbasierte Netzwerke im Finanzsektor’, in J. Esser and E.W. Schamp (eds), *Metropolitane Region in der Vernetzung*. Frankfurt: Campus, 131–53.

Lo, V. (2000) ‘Networking for localised knowledge: The case of the M&A-Sector’, Working Paper Series SFB 403 AB-00-19, Frankfurt M.

Lo, V. and Schamp, E.W. (2001) ‘Finanzplätze auf globalen Märkten – Beispiel Frankfurt/Main’, *Geographische Rundschau* 53(7–8): 26–31.

Martin, R. (1999) ‘The new economic geography of money’, in R. Martin (ed.), *Money and the Space Economy*. Chichester: John Wiley & Sons, 3–27.

O’Brien, R. (1992) *Global financial integration: The end of geography*. London: Pinter.

Pagano, M. (1989) ‘Trading volume and asset liquidity’, *Quarterly Journal of Economics* 104(2): 255–74.

Pagano, M., Randl, O., Roell, A. and Zechner, J. (2001) ‘What makes stock exchanges succeed? Evidence from cross-listing decisions’, *European Economic Review* 45: 770–82.

Porteous, D.J. (1999) ‘The development of financial centres: Location, information externalities and path dependence’, in R. Martin (ed.), *Money and the Space Economy*. Chichester: John Wiley & Sons, 95–114.

Portes, R., Rey, H. and Oh, Y. (2001) 'Information and capital flows: The determinants of transactions in financial assets', *European Economic Review* 45: 783–96.

Santos, T. and Scheinkman, J.A. (2001) 'Competition among exchanges', *Quarterly Journal of Economics* 116(3): 1027–61.

Sassen, S. (1991) *The Global City: New York, London, Tokyo*. Princeton: Princeton University Press.

Schamp, E.W. (1999) 'The system of German financial centres at the crossroads: from national to European scale', in E. Wever (ed.), *Cities in Perspective I. Economy, Planning and the Environment*. Assen: Van Gorcum, 83–98.

Schelling, T.C. (1978) *Micromotives and Macrobehavior*. New York, London: W.W. Norton.

Seifert, W.G. (2000) 'Der Börsenzusammenschluß nutzt der ganzen Volkswirtschaft', *Frankfurter Allgemeine Zeitung* 23, August 2000.

Thrift, N. (1994) 'On the social and cultural determinants of international financial centres: The case of the City of London', in S. Corbridge, R. Martin and N. Thrift (eds) *Money, Power and Space*. Oxford: Blackwell, 327–55.

Veltz, P. (1996) *Mondialisation, villes et territoires. L'âeconomie d'Archipel*. Paris : PUF.

Wahrenburg, M. (2001) 'Die Börse – eine Branche im Umbruch', *Forschung Frankfurt* I/2001, 56–63.

12 Electronic trading and its implications for financial systems*

Helen Allen, John Hawkins and Setsuya Sato

1 Introduction and themes

The adoption of electronic trading systems has transformed the economic landscape of trading venues and is proving a force for change in market architecture and consequential trading possibilities.[1] Electronic trading both removes geographical restraints and allows continuous multilateral interaction (whereas telephone trading allows only the former and floor trading only the latter). It allows much higher volumes of trades to be handled, and in customized ways that until recently would have been technically impossible or prohibitively expensive. This paper considers areas where these enabling effects have been particularly important in wholesale financial markets and how they raise wider policy implications.

The term 'electronic trading' is used in many ways. In this paper, it refers mainly to trading in *wholesale financial markets* (as opposed to e-commerce more generally – see, for example, Long, 2000 for a survey of the latter) and focuses on the central feature of electronic trading systems, automation of trade execution. Such systems usually also feature electronic order routing and dissemination of trade information and may link through to clearing and settlement.

In this introduction, two of the overarching issues that run through this discussion of electronic trading are highlighted. One is the importance of distinguishing between ultimate and intermediate objectives, for example whether a policy concern over transparency is an end in itself or, as argued here, a 'means to an end'. Second is the pervasive presence of network-related policy issues, encapsulated in the question of how, or indeed whether, to respond to perceived problems over market fragmentation. This section begins by presenting an analytical framework for the discussion.

Analytical framework

Much traditional economic literature has little to say about the practicalities of price formation, assuming it as perfect and frictionless or occurring via the fiction

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

of the Walrasian auctioneer. This 'black box' has been illuminated in recent decades as microstructure research has analysed how different market structures and imperfections influence trading outcomes.[2]

This field of research has analysed how market architecture – meaning broadly the key features of market structure such as participation arrangements, venues and trading protocols – influences trading outcomes of prices and quantities. These decisions about architecture also affect aspects of a market's quality – its performance across attributes such as liquidity, trading costs, price efficiency and resilience to shocks. Ultimately, market quality has broader welfare implications – such as through the contribution of the efficiency of the financial system to economic growth and through the performance and resilience of markets to financial stability.[3]

Innovations such as electronic trading can lead to changes in this process. The introduction of new trading technology offers greater possibilities in market architecture, which affects market quality and ultimately influences welfare more widely. Box A characterizes how, across the many interacting and overlapping elements of market architecture and quality, the practical choices have widened, enabling trading arrangements to be varied in ways previously infeasible. Figure 12.1 provides an indicative summary.[4]

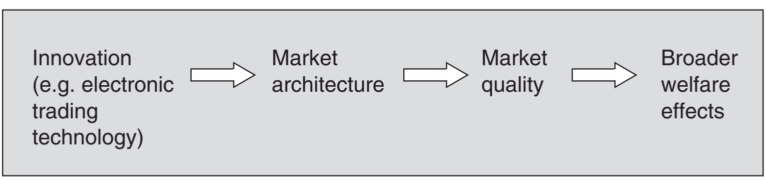


Figure 12.1 Flowchart showing effects of innovation.

Of course, the 'optimal choice' of trading arrangements will mean different things to different parties. For example, owners of trading systems might be trying to maximise market share or income. Traders might be aiming to complete deals at the best prices and at low cost, or with minimum delay or maximum discretion – or indeed, any combination of an enormous range of factors affecting trading decisions. Similarly, there are multiple public policy objectives; for example regulators focus particularly on the integrity of markets and consumer protection while central banks concentrate on systemic financial stability implications.

No single arrangement is preferred by all players and what benefits one set of participants can be disadvantageous to others. The microstructure area is replete with trade-offs, gainers and losers. The multiple objectives at every stage make it unsurprising that there is no unanimity on what constitutes 'optimal' trading arrangements, including at the level of public policy. There is no single, widely agreed 'social welfare' measure to optimize.

Box A – A framework for illustrating the effects of electronic trading

This box shows how innovation in markets – here the introduction of electronic trading – can open up a greater range of possibilities in the trading process, which in turn can affect overall market performance and broader welfare.

Markets can be described in terms of a number of key features which combine to determine the form of the trading that occurs. These aspects of market architecture include which participants have access to the trading platform, the degree of transparency in the trading process, and the trading protocols such as order types and opening hours. Related to these are aspects of broadly defined market quality, such as trading costs and liquidity. In Figure 12.2 they are represented as a circle rather than on two axes as they closely interact and overlap with one another.* They all combine to feed through to 'ultimate objectives', i.e. the effect on broader welfare.

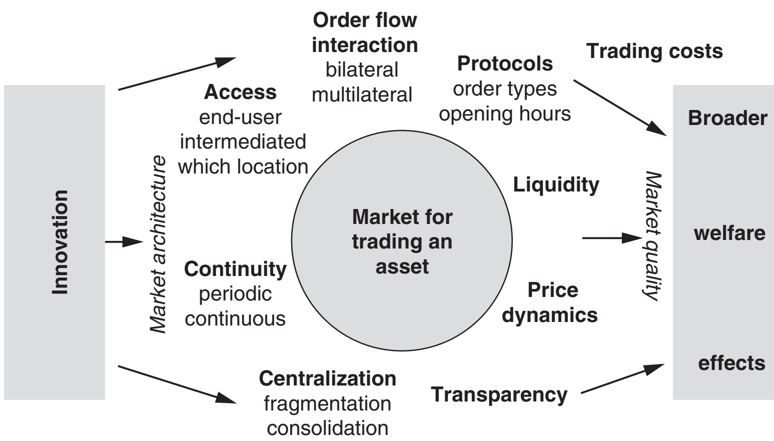


Figure 12.2 Effects of electronic trading.

In terms of Figure 12.2, any key feature of the market (examples of which are **in bold type around the circumference**) may in principle take several, possibly many, forms. For example, access can be for end users or intermediated, geographically limited or unlimited – a theoretical list of possibilities would be very long. (Other indicative examples are placed underneath some of the key features.)

In practice, the trading arrangement for any specific asset in its market segment could be defined by highlighting the relevant items on the lists of key features. For example, in the OTC fixed income markets, access is intermediated by dealers. Transparency is limited, and, so long as this market is telephone-based, there is probably little alternative to these arrangements.

However, the innovation of electronic trading can bring an array of other possibilities within practical reach. In the fixed income example above, order books can become feasible, allowing customers direct access as well as through dealers. Electronic systems can readily disseminate trading information, creating potentially higher transparency.

The result is that electronic trading opens up a far wider range (and combination) of possibilities in the trading process than were hitherto available. This of course begs questions such as 'What to choose? Which combination is best? What are the wider implications?' – issues which are raised in this paper.

* Naturally, any classification of such features is to some degree arbitrary and the above is intended as a presentational device. There are numerous elements that contribute to the overall trading process.

Ends and means – intermediate and ultimate objectives

Which perspective is adopted clearly affects the terms of the debate. The discussion in this paper is directed towards the right-hand side of the diagram in Box A, essentially overall market, financial stability and public interest issues. In this context, it could well be that focusing only on the form of some specific elements towards the left-hand side – be it the details of technology or the specifics of market architecture – could result in a suboptimal outcome from the broader perspective of market quality and social welfare. Indeed, aspects such as market architecture *might more usefully be considered as intermediate objectives rather than as public policy aims in themselves*.

This perspective applies particularly to transparency of information about the trading process (timely information on the prices and quantities of potential and executed orders). The form of disclosure trading systems should adopt is a highly contentious issue. There is debate about how far regulators should, or indeed can, impose transparency rules. Findings about the precise effects of such rules are inconclusive. It is agreed that transparency of trading information is a very important factor in helping markets function effectively, especially for retail interests. However, there is also evidence that after a certain point in certain market segments, insistence on more transparency can be unhelpful and even damaging to market quality (as explained in Section 3). This area, long debated amongst market practitioners, is now recognized in policy discussions.

Regulatory and policy instincts typically – and probably rightly – generally favour more transparency. In many areas of public policy greater openness is widely recognized as beneficial to processes, expectations and outcomes – disclosure practices in accounting and the transparency of the monetary policy process are two important such cases. However, and possibly counter-intuitively, the reverse can be the case in some contexts of trading, where 'too much' transparency can, for example, reduce market liquidity.

In terms of the simple framework above, a policy which makes greater transparency an objective in itself risks ignoring potential negative effects on market quality – and hence on broader investor welfare and the effectiveness of the financial system. Indeed, were transparency to be 'maximized' is a policy end in itself, that aim could prove precisely at odds with the wider objectives to which policy is typically addressed.

The pervasiveness of networks and challenges for policy

Network issues permeate analysis of trading. They vary from the network externalities of markets attracting liquidity to practical consideration of physical network access arrangements. The greater use of electronic systems and linkages for trading highlights these effects.

Although network technologies may broaden access and in principle enable 'more perfect' markets, the immediate reality may prove less benign. For example, the presence of electronic networks can embed existing privileges, with network access and design choices giving strategic advantages to certain classes of participants (See Section 5). And network effects (see Box B) are a powerful influence in this area. They can lead to sustained suboptimal equilibria in market arrangements and a tendency to consolidation. While the latter may bring about significant economies of scale in market processes, if (near) monopoly power emerges there can be undesirable outcomes such as lack of choice and monopoly pricing.

The prospect of such outcomes creates difficult policy questions. First, identifying a suboptimal situation is not easy given the different, possibly conflicting, objectives of affected parties. Then, even where problems are recognized, they may represent a temporary phase of an immature market, which competitive forces and innovation could well resolve. Intervention may be judged inadvisable unless there is demonstrably a sustained problem plus a response which could clearly improve on the market outcome.

Analysis of such situations carries clear dangers of misinterpretation. The multiple, potentially competing, public policy objectives make it likely that not all can be simultaneously achieved. There are dilemmas between intervention or 'wait and see'. The former could risk stifling competition and limiting innovation in fast-changing environments, while the latter might risk missing opportunities to prevent problems becoming widespread.

Responses to market fragmentation illustrate the difficulties, seen especially in equity markets when alternative trading venues become available. Concerns are associated with whether the liquidity of the 'main' market is reduced –

Box B – Network economics effects*

Network economics effects feature strongly in trading systems and help explain commonly observed features of markets – such as consolidation of market liquidity, the advantages experienced by incumbent trading systems, and tipping effects when a market shifts from one centre to another. The underlying economics of these features occur in a number of industries which are structured around a network arrangement – railways and telecommunications are examples.

In these markets, network externalities arise because the value of the network to each participant rises as other participants join. Telephones are a traditional example – in the early days of telephony it was relatively unattractive to join the network since there were few other participants to whom to make calls. However, as the number of subscribers increased, the opportunities for making and receiving calls also increased, enhancing the usefulness and value of the network for all participants, making all users better off.

These positive *network externalities* similarly apply to market liquidity. All other things being equal, it is better to participate in a bigger than a smaller trading network, since each trader brings additional trading opportunities/liquidity. Positive feedback comes about as a liquid market attracts more participants; all participants benefit from the additional liquidity, making the network more attractive to others, and so on.

In the absence of rigidities or other barriers, the presence of these network externalities in a market would imply a *tendency to consolidation*. In the trading context, this would work to bring isolated pools of liquidity together.

However, such consolidation may not occur around an ‘optimal’ system. One reason is *first mover advantage*. An incumbent system may have gained a critical mass of users simply because it was the earliest available. Systems that come to the market later may face formidable hurdles to attract a viable level of participation, even if they offer a better product. Potential users need to believe the costs of switching to the new systems are worthwhile. Moreover, they must expect enough other users will also switch to make the new system an effective, liquid trading venue.

These hurdles may mean users feel ‘locked in’ to a dominant system, in which case a *suboptimal equilibrium can be sustained*. This position can arise whether a system has become dominant through first mover advantage or through consolidation. In the latter case, even if the consolidation occurred around an efficient, technically advanced system, if it comes to be a (near) monopoly the incentives to maintain those advantages can be eroded. The well-known problems of monopoly pricing, technical inefficiencies and abuse of dominant market position may arise.

However, it is by no means inevitable that dominant market positions will be sustained. If an alternative system manages to attract users, it too can enter a virtuous circle of positive feedback. Once a critical level of participation is achieved, the *market can tip* away from the incumbent and towards the alternative. This switch can be abrupt.

Source: The information in this box is drawn from: Shapiro and Varian (1999, Chapter 7), which explains the impact of positive network externalities on industries; Domowitz and Steil (2001a), who analyse how network externalities apply to securities trading; and Economides' (1996, 2001) papers on network economics and finance.

which could mean less market depth to absorb large trades and shocks, reduced price efficiency, higher search costs and price comparison being made more difficult. However, the flip side is that the additional execution routes may reflect greater variety of services, competition to cut costs and innovation in trading systems.

This raises significant questions about the efficacy of policy intervention. It is illustrative of the dilemmas that views can differ starkly even over whether there is a problem to be addressed. If the situation is believed detrimental to public policy interests, there are judgments to be made over whether market forces will resolve difficulties (say, through consolidation following competitive attrition of illiquid venues and/or technological innovations offering linkages). Or whether, and in what manner, to intervene actively.

Outline of rest of paper

The development of electronic trading in wholesale financial markets is considered in Section 2 – where it is observed that the widely differing forms and speeds of penetration can largely be explained by the characteristics of the asset traded and its market segment along with specific factors such as regulation and competitive conditions.

Sections 3 and 4 consider changes to particular aspects of markets' architecture and quality closely associated with the use of electronic trading. In the former, changes brought about in the areas of access, transparency and consolidation/fragmentation are highlighted. The latter section discusses impacts on costs, liquidity and price dynamics, drawing particularly on recent evidence. The lessening influence of physical restrictions on all these areas has brought into stark focus questions of what are appropriate choices, the implications of which are appearing on policy agendas worldwide.

The final section highlights some of these policy issues, emphasizing the perspective of the overall quality of markets and broader welfare. Particular consideration is given to financial system stability issues along with the oversight-related concerns associated with the range and speed of changes in trading technology. As brought out above, it is also notable how network effects in some way influence all of these – from the practical details of system design, to the problems and solutions of oversight, competition and access to network infrastructure, as well the economics of networks which pervade the whole subject.

2 Contrasting development of electronic trading

Looking across financial markets, it is clear that electronic trading has penetrated different sectors very unevenly. This varied pattern of development between types of assets (and market segments within each) can, however, typically be explained by the interaction of a number of factors.

Existing market structures, regulatory and competitive factors and the varied needs of traders have all affected the integration of new technology into mainstream trading. An important element is the asset type, since standardized, homogenous products have proved 'easiest' to migrate to electronic trading. In terms of the schema in Box A, there are clear differences in what possibilities are feasible, both between assets and within different segments of any particular asset market.

The following reviews developments in the equity, fixed income and foreign exchange markets,[5] which between them illustrate the varied development path of electronic trading and its wider effects on markets (the latter being explored in Section 3). Moreover, electronic trading is presumably at only an early stage of development – the future may see, for example, distinctions between market sectors blurring as systems develop for portfolio trading of multiple assets; and advances in networks – perhaps with Internet use more widely integrated. Similarly, the direction of developments to date inevitably reflects the current technology – advances will widen the choices available in next-generation trading systems, potentially enabling further waves of change to market arrangements.

Equity markets: USA versus Europe

Equity markets are the best known, most widely studied examples of electronic trading. The contrasting development patterns in the USA and Europe show how electronic trading can penetrate the market for the same assets in a very *different* manner. Whereas the US equity market has been characterized by a *proliferation* of alternative electronic trading venues alongside relatively few traditional exchanges, Europe has been more notable for the *absence* of separate systems, with electronic trading instead incorporated *within* its many traditional exchanges. In both markets the common features of the liquidity and relative homogeneity of the major equity issues has made it relatively straightforward and cost-effective to introduce electronic trading.

US equity markets

The so-called 'traditional' markets (floor or telephone, albeit with high levels of automation) in the USA are dominated by the three national markets, of which the

New York Stock Exchange and the smaller AMEX are basically order-driven and floor-based. NASDAQ, the second largest venue, is telephone/screen-based (no floor) and has developed essentially as a dealer market with order-driven facilities becoming available more recently.

Separate electronic trading systems have gained a foothold in the USA over recent decades, but it is within the last five years that alternative venues have proliferated. A particular trigger in the US case was regulatory and mainly limited to NASDAQ – the appearance of several ECNs (electronic communication networks) can be directly associated with regulatory changes affecting the display of orders.[6] All these ECNs offer a model based on an electronic order book (which was unavailable in the main markets), though within that they compete on differences in market architecture. Moreover, the USA has significant retail participation in equity markets, business in which a number of ECNs specialize. These country-specific factors illustrate why results from studies of one market may not necessarily generalize to others, a pertinent point given the dominance of studies of the NYSE and NASDAQ in the literature.

European equity markets

In contrast to the USA, most electronic trading facilities in Europe have developed within existing exchanges. Over a period of some years, continuous electronic order books have been incorporated within mainstream exchanges, offering trading methods that in the USA were only available by routing away from the traditional venues.

This has been able to occur in an environment generally less influenced by regulation than equivalent US markets, making European exchanges more open to competitive pressures. Demutualization by many major exchanges additionally gives clearer commercial incentives for innovations and efficiencies. Cross-border competition has led to numerous mergers and alliances among European exchanges, particularly in the last five years. Investment behaviour and trading platforms are moving from being wholly split on national lines to a greater pan-European, sectoral emphasis,[7] particularly with the launch of the euro. It is uncontroversial to predict a further reduction in the number of separate European exchanges,[8] although currently there are several new contenders targeting different segments of the cross-border European market, notably in the market for larger stocks. This juxtaposition of consolidation with proliferation is explored in Section 3.

These developments result in a vastly reduced opportunity, compared to the USA, for separate off-exchange trading systems (especially if designed as an electronic order book) to enter the market – such entrants would presumably have to offer some particular advantage that cannot be obtained on the exchange systems. Bearing this out, there are relatively few separate electronic trading venues operating in European markets.

Fixed income markets

Moves to electronic trading in fixed income markets have been slower than for equities. For many years, bonds of all types were typically traded in telephone dealer markets,[9] into which electronic systems have made (limited) inroads only very recently. Fixed income systems proliferated in the late 1990s but are now consolidating – an annual survey by the Bond Market Association (2001) identified almost 70 systems in 2000, up from only 11 three years earlier, although by 2001 only 49 were still in operation.

The variety of automated trading system types already introduced is notable (a variety previously more characteristic of equity markets), given the uniformity of traditional bond market arrangements. There are now order-driven markets in addition to several automated versions of dealer markets, offering a range of participation and access arrangements for dealer, inter-dealer and customer sectors (for a taxonomy, see BMA, 2000).

The later arrival of electronic trading in fixed income markets compared to equities *reflects distinct differences between the two*. Fixed income products are far less homogenous, with many more separate and individually less liquid issues than equities, making it technically more difficult and more expensive to introduce automated systems. *The Economist* (2000) suggests there are over 4 million fixed income securities on issue in the USA (varying in coupon, maturity, frequency of interest payments, etc.) compared with a few thousand listed shares. Trading style differs – relative to equity markets there tend to be fewer but larger trades,[10] with many participants holding issues to maturity. Moreover, while it seems likely that the trading of equities on centralized exchanges encouraged early transitions to automation, the opposite position delayed bonds from making the move.

Within the fixed income sector, electronic trading has made most inroads into certain *government bond markets*. It is estimated that 40 per cent of US Treasury securities transactions in 2000 were done electronically, nearly double the volume a year earlier (Moszkowski *et al*., 2001). Similarly, in Italy about half the trades in government debt are now conducted on the MTS system developed with the cooperation of Italian public authorities.[11] Electronic trading in corporate bond markets remains much lower (only 10 per cent of US corporate bond volumes in 2000 according to Moszkowski *et al*.) as they are far more heterogenous. A common pattern has been for platforms to begin trading government bonds, later expanding into other, more heterogeneous, fixed income issues. Though this reflects the greater amenability of the relatively more standard and liquid government securities to current electronic trading platforms, this distinction may reduce as electronic systems develop to accommodate trading of less liquid issues – see the discussion in Section 4.[12]

Foreign exchange markets

Electronic trading has had an important presence in the inter-dealer spot foreign exchange market for over a decade; the BIS triennial survey shows that the majority of interbank trading in the major currencies is now conducted electronically. For some years there have been two major systems (EBS and Reuters) which now each specialize in particular major currency pairs. The development of these systems is described in Chaboud and Weinberg (2002). Both systems have been designed as order books, in which dealers can see the best bid and offer in the market, alongside the best bid and offer that they could trade subject to their institutional credit limit structure.

These electronic systems are now used for the majority of spot inter-dealer trading in major currency pairs. Consequently, the inter-dealer segment of the market has mostly (but not entirely) moved away from voice broking, and the electronic systems now act as a standard reference for pricing. While the structure of the foreign exchange market before the introduction of electronic trading was (rather akin to fixed income) a fragmented bilateral telephone market, the rapid adoption of systems in the inter-dealer sphere reflects the *liquid, homogenous nature of the product* that can be traded in standardized units. The latter points presumably explain the earlier presence of electronic trading in foreign exchange compared to fixed income markets.

While electronic trading has come to dominate the inter-dealer market, systems have made far less impact on the business of large corporate customers. However, this may be on the point of changing as several Internet-based systems aimed at this area are now being rolled out. These systems promise more flexibility (e.g. tailored quantities and currency pairs available) and utilize the Internet's ability to connect disparate and distant parties at low cost.

3 Effects of electronic trading on market architecture

This section considers effects on market architecture with which electronic trading is closely associated – focusing on consolidation/fragmentation of markets, decisions about the type of participants' market access and the form of transparency in the trading process. Substantial changes are taking place in these areas in the wake of electronic trading lessening many previous physical rigidities. As illustrated above, these developments in market structures are closely related to the characteristics of each market/asset type.

Fragmentation and consolidation

Electronic trading is now closely associated with many of the important pressures affecting ebbs and flows in the number of trading venues.[13] On the one hand, it can stimulate a proliferation of venues, by making markets more contestable as a result of them being cheaper to enter and enabling a greater variety of products and specialization of trading services. On the other hand, electronic systems can enable linkages to bring together sources of liquidity and to harness efficiencies that contribute to consolidation. And affecting all these, the speed of development of the facilitating information network technology has led to a more rapid pace of change.

These effects are of course a feature of any dynamic market, where opportunities for new entrants result in a tendency to fragmentation, typically followed by some reversal when not all attract sufficient business to operate on a profitable scale.[14] The powerful influence of network effects in this area (see Box B) means that a proliferation of similar trading systems which individually attract little liquidity would be expected to be a transitory phenomenon. Only those both adding some value and – crucially – attracting, retaining or linking to a sufficient amount of liquidity, will survive.

While these effects imply a tendency to consolidation over time, in any particular market the extent to which electronic trading is currently motivating a phase of fragmentation or consolidation is very dependent on existing structures. In the previously OTC fixed income market, new trading systems are starting to bring together larger groups of users, consolidating sections of the market which formerly relied on bilateral communication. Similarly, users of the main foreign exchange systems now transact through a centralized market, in contrast to fragmented telephone arrangements. By contrast, in equity markets, which were typically dominated by centralized exchanges, the very presence of distinct alternative trading venues (along with large banks and brokers increasingly able to internalize their order flow by offsetting one transaction against another) can increase fragmentation. Yet in some cases the overall balance may be hard to call – consolidation and fragmentation effects can operate in parallel. For example, European equity markets have also seen numerous recent efforts at mergers, alliances and linkages of various kinds.

This ambiguity affecting judgments on the degree of fragmentation/consolidation in any market's trading is added to by fragmentation taking many forms. As discussed in Lee (1998), there is no single dimension by which to evaluate fragmentation. To illustrate, even though the number of separate exchanges (or similar marketplaces) may fall, the range of systems through which orders can be placed may grow if venues offer a wide choice of order routing. For example, an exchange may offer a main market, an upstairs arrangement for block trades and onward routing to associated markets. Alternatively, if there are linkages giving traders seamless access to a range of markets, it may matter little how many underlying venues exist. These effects, making judgments about fragmentation/consolidation far from straightforward, are amplified as the range of technological possibilities increases.

Moreover, and in an innovative market, the degree of concern about fragmentation may alter quickly. Even in the relatively short history of electronic trading, there are cases of fragmentation effects being overridden by technology. For example, systems can now offer 'virtual consolidation' on a single screen to combine information from stocks traded at multiple venues. Or solutions may take the form of 'smart search agents', automated devices to seek out disparate sources of liquidity (see Section 4).

From the perspective of market users and policy makers, the important question is whether the available routes can bring together liquidity in an appropriate manner to meet trading needs and enhance market quality. For example, whether there is sufficient order interaction to allow effective price formation and depth of liquidity, to the benefit of the ultimate end users. Ways in which electronic trading is influencing these matters are discussed in Section 4. However, as brought out in the introduction to this paper and discussed further in Section 5, it can be difficult even to gain consensus on whether there is a problem of fragmentation, and still more difficult to judge whether a policy response is appropriate, and if so what should be the response.

Market access

Forms of access

Electronic trading can widen access to trading systems across several dimensions. Physical limitations that once rationed access to traditional venues no longer bite, meaning additional users can now participate at minimal marginal cost, removing the economic need to limit access through membership restrictions. At the same time remote linkages remove geographic limitations on the pool of potential users, and continuous multilateral interaction is enabled. From the system providers' viewpoint, the opportunities to enter the market are also greater than hitherto in the light of the fall in costs and enabling technology. And from the perspective of the issuer of securities, there can be opportunity to access a wider pool of potential investors.[15]

This turnaround in the economics of access means that in principle arrangements can be decided more in response to the needs of the market. For example, whether there is a role for intermediaries and whether it is appropriate for different customer types (e.g. retail/wholesale or by different institutional status) to participate in the market can become more a matter of choice rather than being effectively dictated by external constraints. Changes in these areas can affect the whole design of a market.

This is not to say that limitations on access to markets no longer exist – in practice they still have strong effects, albeit with different forms proving more important. Notably, the legal/regulatory factors associated with cross-border transactions now appear to be the more relevant geographic constraint.

Additionally, there may be differential access to telecommunications infrastructure networks – termed 'access asymmetries' by Bar (2001). For example, dedicated private lines or the use of the Internet influence which classes of participant can have physical access to systems.[16] Suboptimal outcomes could arise, for example if intermediaries own new trading platforms that in other respects would disintermediate them – their incentive may be to ensure the network design retains their privileged access positions. Implications of the risks of anticompetitive biases in networks are discussed in Section 5.

Access and intermediation

The greater access possibilities offered by electronic trading have perhaps most obviously brought into question the role of intermediaries. There has been some shift away from pure dealer structures towards continuous auction arrangements where users can transact directly with one another. This has especially occurred in large, liquid markets (notably major equity and foreign exchange) where end investors are likely to be able to match their requirements directly over a reasonable period of time. This wider access to trading systems increases competitive pressures on dealers and typically forces those which remain to focus more on value added services such as corporate finance, advisory services and risk management. However, in OTC markets, where a larger portion of trades are characterized by asynchronous supply and demand in less liquid securities, the matching services of a dealer have a clearer role and this is reflected in many electronic trading systems in such markets incorporating dealer structures, as illustrated in Section 2.

However, despite the developments (such as customer access to order books) that might have been expected to marginalize much of their role, dealer intermediation remains. Reviewing literature pertinent to this issue, Madhavan (2000b) highlights some features, such as the significant costs an end user might face in order to monitor changing market conditions, that could explain their continued importance. Supporting this, anecdotal evidence collected by CGFS (2001a) from customers in foreign exchange and fixed income markets pointed to some valuing the personalized research, advice and execution offered by dealers. Moreover, in practice, systems are some distance from being able to seek out all sources of liquidity automatically, and disparate sources of liquidity (e.g. due to upstairs arrangements) make this search a valued service of intermediaries.

Transparency

The uneven effects of changes in transparency

One of the commonly cited benefits of electronic trading is that it can facilitate greater pre- and post-trade transparency.[17] While this is undoubtedly true, there are significant questions, both theoretical and practical, about the extent to which implementing greater transparency across all market segments benefits the quality of a market. Underlying the issue is that transparency arrangements affect the balance of information among participants. Evidence from a range of studies (see Madhavan, 2000b) demonstrates that this influences the degree of information in the order flow, price discovery and liquidity.[18]

Changes to transparency rules tend to benefit one group of participants and their objectives at the expense of another, creating winners and losers. While in many respects the literature on the effects of transparency is inconclusive (see, for example, Ganley *et al.*, 1998), the presence of these trade-offs is very evident. This highlights the importance – stressed in the introduction to this paper – of the perspective of the debate. From the public policy viewpoint, recognition that transparency arrangements can have uneven influence on ‘ultimate’ objectives such as market quality and broader welfare is clearly crucial. Even though understanding of this area is very incomplete, the evidence that there is not a simple, unidirectional relationship between transparency and quality of markets deserves considerable weight in policy making.

Some flavour of why greater disclosure may not necessarily benefit the overall market is given in the following stylized examples. An illustration of a tension between *post-trade transparency* and liquidity occurs in a multiple dealer setting such as in many government bond markets. Faced with an unpredictable flow of large customer orders, dealers who maintain a continuous presence in the market seek to manage the risks arising from sharp variations in their inventory of securities by inter-dealer trading to rebalance their holdings. Were stricter post-trade transparency imposed in terms of requiring more rapid publication of large transactions, it would reduce dealers' opportunity to conduct this inventory adjustment. This could increase their risk management costs – which may be passed onto customers – and could lead to a less efficient allocation of risks in the market. Both liquidity and price discovery could be impeded. (Gravelle, 2002, provides more details on the above arguments.)

Equivalent tensions with *pre-trade transparency* requirements can arise where transactions contain (and are motivated by) private information reflecting legitimate investor research/beliefs or portfolio strategy. Were disclosure imposed which revealed 'too much' about intended trades, it could effectively expropriate that private information for the public trading venue. The predictable result of such rules would be that traders would act to minimize the cost of the loss, for example by splitting the trade to reduce the observable information content or by switching venues to avoid the regime. Or they might exit the market entirely.

Transparency – potential and practice with electronic trading

Electronic trading creates the potential for a very high degree of transparency across the whole trading process. In principle, systems can disseminate real-time pre- and post-trade information marketwide. Conversely, they can operate with minimal information leakage, in a manner that trading based on personal contact could not achieve. As electronic systems become more sophisticated, they make it more feasible to move along the multidimensional spectrum of transparency. They can more readily meet different users' preferences regarding information, some of which were once too complex to put into practice.

For example, the basic demand for anonymous trading is now met through many electronic systems. Some of the systems becoming available are specifically designed to eliminate (pre-trade) information leakage, enabling users to specify precise orders without giving away potentially valuable information to competitors.[19] Other systems offer choices such as 'iceberg' orders that are automatically matched if hit but are not visible on an order book. In contrast, the upsurge of electronic order books in equity markets has often led to the arrangements being more transparent than the structures they replaced. Presumably the practical possibilities will grow as the technology develops further.[20]

In itself, this existing variation in transparency arrangements suggests that the form and degree of disclosure vary with largely market-specific factors. Considerations include the perceived role of the information in attracting liquidity to the system, the needs of its range of users and style of trades (e.g. retail/wholesale, whether market-moving) and the commercial value of the data. And importantly, different classes of trading systems lend themselves to different forms of transparency: the style of information concerning a call market differs from that readily available from an order book or a dealer arrangement. Regulatory requirements may or may not constrain the actual outcome, given that there are strong reasons why systems typically would choose to make available some degree of trade information (e.g. to attract liquidity and/or for commercial data promulgation).

Some likely effects of electronic trading on transparency

Segregation of trading arrangements largely according to transparency regime has long been a feature of markets. Notably, virtually all exchanges have particular arrangements for block trades ('upstairs trading'), with lower transparency requirements, often in the form of delayed publication. The integration of electronic trading hugely increases possibilities for different arrangements, in principle virtually anywhere on a spectrum between complete transparency and complete opacity. An outcome may well be an even greater variety in transparency levels across (and within) different trading venues than seen hitherto.

Moreover, if electronic trading enables the implementation of systems which give 'appropriate' incentives in trading behaviour (such as to input 'truthful' orders, as explained in note 19), one result could be greater efficiency of price formation. As set out in the introduction, this argues for seeing transparency as a means to an end for aiding market quality, rather than as an end in itself.

The current regulatory focus on transparency in securities markets largely reflects concerns that greater choice of trading venues/routings raises about level playing fields and fairness of information across the whole market. For the context of the immediate discussion, there are clear risks that a response involving an attempt to impose common transparency standards could be counterproductive if it overlooks the wider repercussions of disclosure rules and creates distortions by neglecting differences between/within markets. These questions are particularly important across wholesale, professional markets. For retail (i.e. non-market-moving and non-information-carrying trades – albeit not the focus of this paper), the concerns are very different, relating more to consumer protection and informed consumer choice, where the benefits of greater transparency seem less ambiguous.

4 Effects of electronic trading on market quality

Closely related to market architecture are aspects of market performance. This section looks at influences of electronic trading on transactions costs, price dynamics and liquidity, all of which contribute to the overall quality of markets.

Trading costs

At the outset of this paper, the performance of financial systems was recognized as having broader welfare effects. Trading costs are one of the more direct indicators of this performance. For example, in a recent study of the wider implications of trading costs, their reduction was shown to be associated with a lower cost of equity capital, which has macroeconomic significance – see Domowitz and Steil (2001b).

Benefits such as these could be assuming more importance since electronic trading and associated computing advances have given new impetus to trading cost reduction across all fronts. There is now scope to reduce what was once a 'set cost' of business – one reason for the greater focus by institutions on analysing and cutting trading costs. These costs for end users can be divided into explicit costs (e.g. physical overheads, fees, commissions, taxes, costs of clearing and settlement) and two main types of implicit costs – bid–ask spreads and market impact costs. Electronic trading is offering routes to reduce all of these.[21]

Looking initially at *total trading costs*, empirical studies find these are lower for institutional investors on automated rather than traditional markets, even after controlling for the different mix of shares traded in the two types of market. For example, Domowitz and Steil (2001b) suggest total cost savings in the USA of around 30 per cent from using automated systems. A study of over half a million institutional equity trades by Domowitz *et al*. (2001) – summarized in Table 10.1 of Degryse and van Achter (2001) – suggests that costs have been falling over time and that they tend to be lower in markets where automated trading dominates.

Explicit costs

Electronic systems involve lower set-up costs than trading floors, especially when an existing system can be adapted for a new product. More significantly, by replacing labour-intensive processes, they markedly reduce operating costs; Domowitz and Steil (2001a) show such reductions have typically been 50–75 per cent.

The cost savings from shifting to electronic systems will differ for different users. A more competitive market reduces the ability of exchanges to cross-subsidize different types of trading activity, meaning that the allocation of costs may shift under electronic trading.

In addition, electronic trading can further lower costs if integrated into straight-through processing (STP) arrangements, which allow trades to pass automatically through to final settlement without further manual intervention. As well as the savings due to automation and error reduction, costs can be cut by linking the execution, clearing and settlement of trades to the procedures for controlling market and operational risks. According to CGFS (2001a), some institutional investors believe that the prospect of STP could be the greatest single potential benefit of electronic trading.

Implicit costs

The *bid–ask spread* paid by users of dealer markets could be regarded either as a payment to the dealer for providing liquidity or as a rent charged for accessing the market. It must cover dealers' normal costs of doing business (e.g. order processing, inventory costs, overheads) and a risk premium to compensate for the losses incurred when trading with well-informed traders (adverse selection). Electronic trading may reduce all these components. It obviously allows cheaper order processing and lower overheads in general. Moreover, electronic trading's potential to make markets more transparent in conjunction with anonymity reduces the risk premium and may drive down profit margins, by increasing competition between dealers and aiding price comparison. Some auction markets give end investors direct access, which might be regarded as them earning the bid–ask spread by providing liquidity. Profit margins are likely to be particularly low in electronic order books that allow access to end investors.

A number of empirical studies show that realized bid–ask spreads on electronic systems are similar to, or lower than, those on floor- or telephone-based systems.[22] German Bund-futures provided a case study when in the mid-1990s they were traded in large volumes under otherwise similar conditions on both the electronic Eurex market in Frankfurt and the LIFFE floor in London. Comparisons by Breedon and Holland (1998) and Frino, McInish and Toner (1998) found that spreads were generally wider on the floor exchange, at least when the two exchanges had similar volumes.

Market impact costs refer to any adverse impact on price as a result of information associated with the trade leaking ahead of execution, or because the trade is large enough to affect significantly supply and demand in the market or signal a predictable trade to come. As discussed above, electronic trading may reduce these market impact costs, for example through the use of the pre-trade non-transparent systems. This is an area where electronic trading is widely held to have great potential – with several new systems aiming at adding value. More generally, studies by Vila and Sandman (1995) and Pirrong (1996) find that prices are less sensitive to volumes in automated than traditional markets.

Price dynamics

Effective price discovery is important beyond the immediate asset market, since it underlies the accuracy of price signals to agents in the wider economy. This enables appropriate investment (and other) decisions which have widespread economic impact, contributing to ultimate objectives of broader welfare.

Price formation in electronic trading systems is the outcome of precise order execution algorithms, in contrast to the trading floor or phone-based systems where relationships may matter as much as price or size.[23] Furthermore, electronic trading is allowing basic algorithms to be extended to better meet trading needs and, for example, some now permit very detailed trading plans with contingent orders reflecting the various nuances of preferences. Systems are being developed for 'near matches': when bids and offers do not coincide, a computerized negotiating system may try to reach a compromise price.

In general, electronic processing should allow orders to reach the central market faster because of higher processing speeds than with manual processes. Prices should therefore incorporate information more quickly. Domowitz and Steil

(2001a) found that most empirical studies show that electronic systems are more efficient in this sense than traditional trading venues. However, in some studies the difference is quite marked while in others it is rather small.

Effects on price volatility

Studies comparing observed price volatility in traditional and electronic markets generally find that volatility is less or about the same in electronic markets. The following recent studies give a flavour of the findings. Jiang, Tang and Law (2002) observe that price volatility dropped in the Hong Kong futures market after the introduction of electronic trading but as a similar drop occurred in the spot market, their results suggest that electronic trading itself had little effect on volatility. Chaboud and Weinberg (2002) report mixed evidence for the foreign exchange market. Using over a century of data from the London Stock Exchange, Green, Maggioni and Murinde (2000) find that, other things being equal, lower transactions costs reduce price volatility. Madhavan (1996) suggests that in large liquid markets price volatility is lower when there is greater transparency. These latter results would imply that more widespread electronic trading should lower price volatility.

Some market participants suggest that price volatility is more ‘visible’ as price transparency has increased in some markets – cited by CGFS (2001a). Systems are technically capable of extremely rapid change (e.g. EuroMTS can absorb 150 price changes per second) and in some markets the use of pricing engines allows orders to be generated more rapidly than hitherto. The micro design of trading systems may also affect price volatility over very short periods. For example, Soejima (2001) shows how changes to the system of execution in the Japanese futures market had to be amended when it was found that traders had insufficient time to respond to new information on the order book.

Effects on price discovery

The effects of the increased opportunities for trades to be executed with little interaction with other orders and/or without reaching a main market are keenly debated. Electronic trading is closely associated with the issue since it enables many systems to perform this order matching, but the debate – which is typically voiced with respect to the equity market – goes much broader.[24]

Three factors have a particular influence. The first is where orders become more thinly spread among fragmented venues. The second is increasing internalization, where brokers match buy and sell orders internally from their own customers at prices (broadly) determined on another market, sending only the net balance to the exchange. The third is the number of formal, non-exchange ‘crossing’ systems for institutional trades that are now also part of the market, which match at the prevailing main exchange price.[25] There seems to be no clear conclusion about the combined effect of these on price formation, despite their increasing importance in markets, and Hendershott and Mendelson’s (2000) formal modelling shows ambiguous effects on market quality. The following comments give a flavour of the arguments.

One issue is how crossing or internalization effectively 'free rides' on central price discovery and hence the extent to which price information is a public good. For example, Lin, Geng and Whinston (2001) describe crossing as a 'parasite of real exchanges' while Picot, Bortenlaenger and Roehrl (1995) argue it impinges on the 'property rights on the price discovery' of an exchange. There are also regulatory concerns including whether internalized transactions are conducted at a 'fair' price and whether crossing systems encourage attempts to manipulate the main market just ahead of the price taking for the crossing trades. Transparency can also be reduced by internalization or crossing, though this depends on the trade publication rules prevailing in any centre.

However, significant cost savings are available[26] from crossing and internalization – although the extent to which customers benefit from this can vary. Additionally specialist crossing systems are used by institutions to avoid information leakage. Both these may have some bearing on pricing.

Another aspect of the question is whether there is a 'genuine' problem regarding pricing. Other markets (for example, many commodity markets) operate on the basis of the 'central' market reflecting more the net balance of buy and sell orders rather than gross volumes. The resultant price should be the same, as all the supplies and demands are still contributing to price formation. It could be argued that this form of price determination is more cost-effective and hence generally beneficial to broader welfare provided the regulatory safeguards operate satisfactorily, such as ensuring customers have the means to access a fair price.

The policy questions closely relate to those surrounding the problems of fragmented markets and of transparency. Problems might be expected to be self-corrected if arbitrage processes, especially in combination with more effective information systems and market linkages, lead to the natural migration of business to platforms where there is confidence in the price formation and integrity of the market. On the other hand, there is scope for self-correction to be limited by rigidities (such as the dominance of some cartel or private incentives of intermediaries) while it may also be held damaging to allow even 'temporary' problems to continue (e.g. the practical problems of thin, illiquid markets).

Liquidity

Liquidity is essential for trading systems. It enhances the effectiveness of the market overall, reducing costs by narrowing spreads and giving depth such that prices are less affected by particular trades. Liquid markets are typically better placed to absorb shocks than less liquid ones, contributing to the robustness of financial systems. Moreover, as discussed in the previous section, liquidity is an essential ingredient of price discovery and hence price signals for the wider economy.[27]

Despite the technological and strategic efforts of electronic trading systems to attract liquidity (see below), it is generally those with existing access to order flow (particularly within an existing exchange) that have experienced viable volumes. Few 'stand-alone' systems have achieved this. This, however, probably reflects the powerful network effects of liquidity as much as the characteristics of the electronic systems themselves – with liquidity attracting (and 'locking in') liquidity. There are converse hurdles to attracting business to new venues in the absence of many other users, even if a technically 'better product' is on offer (see Box B).[28]

Nevertheless, in a world of electronic trading, liquidity is much more mobile. Orders can be rerouted to the preferred system and to the best prices, and franchises can be quickly lost. This 'tipping effect' was seen when Eurex, within around six months in early 1998, took all the volume in the futures on the ten-year German Bund contract from the previously dominant LIFFE floor.

Electronic trading systems' ability to build liquidity

Focusing on the level of the trading system design,[29] electronic systems are developing a number of ways to attract liquidity and seek it out from disparate sources. For example, systems can link to institutions' order management systems to interrogate potential orders on their blotters of prospective trades to seek out possible crossing matches. In due course, traders may be able to use 'smart agents' to search across systems to locate disparate sources of liquidity.

Electronic trading systems may encourage issuers to standardize their offerings (which can concentrate liquidity), particularly in more heterogeneous securities such as fixed income. There has been some suggestion of this in government bond markets with the tendency to reduce the number of separate issues. Minimum size limits on issues to be eligible for certain trading systems similarly encourage issues to be made in larger size or reopened to maximize liquidity.

Though electronic trading has typically come later to less liquid assets (as illustrated in Section 2), routes are being developed to enable their cost-effective trading. Here, the most effective automation solutions may not be those that reproduce traditional procedures. For example, in fixed income markets, there is an expectation that small, less liquid issues could get swept up into automated trading of portfolios which offer certain characteristics, rather than being traded bilaterally on their own merits. This kind of development offers the potential to garner individually disparate, illiquid securities and pull them into a larger liquidity pool – for example, see the discussion in Fan, Stallaert and Whinston (2000).

Another way in which electronic systems are enabling the more effective trading of illiquid securities is by reviving the use of periodic call auctions (explored in Schwartz, 2001). A number of stock markets now trade less liquid securities in call auctions, commonly one to three times a day, concentrating liquidity that otherwise would have been thinly dispersed across a longer period. Such securities may not trade effectively in the continuous auctions that typically form main markets – for example, Steil (2001) describes how the Warsaw stock exchange, re-established in 1991, initially traded stocks in a weekly call, moving to daily calls and later (for some stocks) to continuous trading as volumes grew to give sufficient liquidity.

5 Concluding remarks: wider implications and policy

This final section does two things. First, it draws together some wider implications of the effects of electronic trading in three areas: the regulatory framework/oversight issues; networks and competition; and financial stability. Second, in each of these areas it draws some tentative policy conclusions. As already set out, there are multiple, possibly competing, public policy objectives along with uncertainties about the net effect of changes in markets and their transmission to broader welfare. The recognition of these uncertainties and ambiguities in itself deserves considerable weight in policy making.

Regulatory framework and oversight issues for markets

Developments associated with electronic trading test longstanding institutional structures and appear on regulatory agendas worldwide.[30] Notable issues include:

- frameworks for regulation: notably whether to (continue to) differentiate the institutional status and oversight regimes applying to exchanges and non-exchange trading systems;
- the appropriate level of detail for official involvement in microstructure matters – for example, whether transparency rules are required or can be enforced, and in what degree of detail; or whether fragmentation of markets requires an active response; and
- cross-border issues raised by remote access to trading, including whether countries' different regulatory regimes lead to problems caused by regulatory arbitrage; and clarification of jurisdiction over legal and regulatory arrangements.

It is worth reiterating that in immature, fast-changing markets, some perceived problems may be temporary effects perhaps associated with early stages of product cycles, whereas others may reflect sustained suboptimal positions. The questions associated with fragmentation illustrate this. Policy treads a difficult line between imposing requirements that restrict innovations, while maintaining market integrity and confidence in periods of rapid change. Even if aspects of markets are perceived as functioning suboptimally, it is recognized that specific intervention can be counterproductive. Greenspan (2000) captured the concerns, stating that the authorities 'would do well to heed the advice offered to the medical profession and, first, do no harm'. Since market structures that had given rise to concerns may change rapidly, the presumption may be that specific intervention is inadvisable unless there is demonstrably a sustained problem.

Where action is deemed preferable (say, where correcting market forces are believed to be weak), there will also be differing regulatory stances about solutions. These could range from ensuring facilitating frameworks such as clear legal codes, action on competition policy such as removing restrictive practices, to specific micro rule-making on, say, trading protocols. When weighing choices, the differences between and within markets are important – as already evidenced by the varied adoption of electronic trading and the variety of services becoming available. Added to this is recognition of the imprecise understanding of the net effects of changes in market structures/rules – such as the mixed impact of transparency rules. All this supports the avoidance of a 'one size fits all' approach and argues in favour of being wary of imposing detailed, cross-market rules at a high level.

Electronic trading systems still have further to go in capturing the subtleties of trading – and a similar comment applies to market structure research and its application to policymaking. However, while technology's effects raise many difficult issues in markets, technology can also contribute to solutions:

- It can be applied to enable more efficient and effective collation of data on market performance and behaviours – helping both oversight and understanding of markets. For example, the fulfilment of market-maker obligations could be monitored automatically, or erratic market movements, whether due to trader errors or more fundamental reasons, could be identified rapidly.
- It may directly offer solutions to problems, such as the means to build information systems or link fragmented pools of liquidity.
- It can help participants make better informed decisions, for example by enabling more appropriate transparency arrangements, providing greater information about order routing and features of the assets.

Networks and competition

The raised profile of 'physical' (or indeed 'virtual') network issues in trading comes in addition to the better known, powerful network effects. Some associated policy issues, such as fairness of network access, may be akin to those faced by regulators of other industries such as telecoms where the networks have been opened to competition. From the interest of financial authorities, there are also parallels with oversight of payment systems, in which, quoting from Bank of England (2000), 'as with other significant components of the economic infrastructure, there is a public policy interest in ensuring that a competitive environment exists and that any competitive abuses are curbed'.

As highlighted in Box B, the competitive landscape can shift rapidly. For example, outcomes of competing but inter-operable systems are likely to become more common in securities trading (as seen in, for example, telephone networks). This occurs as network inter-operability and linkages between systems contribute to the lessening of physical hurdles to entering markets, for example removing the disadvantage of requiring separate screens for alternative systems. Moreover, while it is never easy to compete successfully with a dominant system, the effects of electronic markets open up more possibilities for tipping to occur, a prospect reinforced by the lower cost base of electronic networks.

Overall, while electronic linkages might in principle open up markets to competition and enable more 'level playing fields', it is not the technology but its implementation that determines the fairness of arrangements. As described by Bar (2001), architectural biases can exist in electronic network markets at least as strongly as, and possibly less visibly than, in traditional structures. Networks can be designed with strategic advantage to certain players, and anticompetitive practices (which restrict the scope for self-correction) may be hard to recognize if they are embedded in the system protocols themselves.[31] However, electronic network architecture should in principle be more adaptable than traditional market structures, implying that, where identified, problems may be easier and speedier to address.

Financial system stability

Many financial stability issues closely associated with electronic trading are common to discussions of technological innovation and e-finance more generally. The magnitude of uncertainties in conjunction with potential for rapid, wide-ranging change in themselves motivate interest in the whole area.[32]

The concerns typically are due to shifts in emphases rather than completely new risks. For example, the impact of operational problems can be greater in a technologically dependent market. Moreover, and as discussed in Turner (2001), potential systemic implications can arise from the involvement of new and different firms in financial markets, making it more difficult to monitor the linkages and assess the risks to which sectors may be exposed. Especially, the increased importance of non-financial firms, such as telecoms and IT companies, in processes could make channels through which systemic threats arise harder to identify and anticipate.

Many recognized effects of electronic trading – which include opportunities to harness efficiency gains, better market information, handle higher volumes and lessen physical constraints on trading practices and participation – have a role in contributing to the adaptability and stability of the financial system. And while, as with the expansion of any new market, the route will doubtless end up littered with underperforming and failed systems – this in itself does not necessarily carry systemic threats. Indeed, if this brings about greater strength in the remaining platforms, it should contribute to financial stability.

The ambiguity of effects on market performance remains striking. Most observations highly relevant to policy remain in some way equivocal. For example, liquid markets are more resilient than illiquid ones; yet electronic trading is credited with both fragmenting liquidity and enabling it to be brought together from disparate sources. Nevertheless, the potentially volatile nature of liquidity and markets in general, combined with the speed and unpredictability of technical developments, remain in themselves reasons for vigilance.

While electronic trading has brought a range of policy issues to the fore, the associated technological advances may offer routes to solutions. This paper has not attempted to predict outcomes. But it seems likely that the direction of resolution of many of the current questions may lie with the technology itself. And likely too that, as with the assimilation of previous technologies, 'electronic trading' will before long cease to be considered as a distinct issue.

Glossary

The literature on e-finance has spawned a lot of technical terms. These are the meanings assigned to terms in this paper. Other papers may use some terms in other ways.

Terms in *italics* in references denote a cross reference.

Algorithm (order execution) Rules to determine the method by which orders are matched.

AMEX A major US equity market, organised as a floor-based *order book*.

Anonymity Non-disclosure of identities of counterparties.

Architecture (market) Broadly, key features of market structure such as participation arrangements, venues, trading *protocols*.

Bid–ask spread Difference between the buying and selling price.

Block trade Large, potentially market-moving trades, which may be exempt from an exchange's usual disclosure rules.

Broker Firm which operates in a market on behalf of other participants to arrange transactions without being a party to the transactions itself (compare *dealer*).

Call market An order-driven market where orders are processed at discrete points in time, being matched by a particular *algorithm*. (Referred to as *crossing system* where prices are matched at the price in another market.)

Clearing The process of transmitting, reconciling and sometimes confirming instructions to transfer instruments prior to *settlement*.

Consolidation of trading Bringing together aspects of the trading process (*order routing*, *execution*) into a smaller number of markets (compare *fragmentation of trading*).

Continuous market Market that trades on an ongoing basis (compare *call market*).

Crossing system System matching buy and sell orders at a price determined in another market.

Dealer Firm whose primary business is entering into transactions on both sides of wholesale financial markets and seeking profits by taking risks in these markets (compare *broker*).

Demutualization Process of exchanges converting from a cooperative organization owned by member *dealers* and *brokers* to a profit-oriented company.

ECNs Electronic Communication Networks, trading systems used for US equities, notably those listed on *NASDAQ*.

E-commerce Sale of goods and services carried out over the *Internet* or other public networks.

Economies of scale Situation where unit costs drop as volume increases.

Electronic trading Use of electronic means for transforming orders into trades.

Eurex A Swiss/German electronic derivatives exchange offering trading, *clearing* and *settlement* on one platform.

EuroMTS A trading platform for euro-denominated government bond benchmarks of nine euro area countries.

Execution Turning matched orders or trade proposals into actual trades.

Explicit trading costs Costs such as market access fees, commissions, tax, *clearing* and *settlement* costs and staff and information technology overheads (compare *implicit trading costs*).

First mover advantage Possible ability of first entrant to a market to achieve a dominant position, e.g. by setting standards or through establishing a dominant brand name.

Fragmentation of trading Division of some aspects of the trading process (*order routing*, *execution*) between different markets (compare *consolidation*).

Iceberg order An order (partially) invisible on the *order book*.

Implicit trading cost The *bid–ask spread* and *market impact cost* of a trade (compare *explicit trading costs*).

Internet An open worldwide communication infrastructure consisting of interconnected computer networks and allowing access to remote information and the exchange of information between computers.

LIFFE London International Financial Futures Exchange.

Liquidity Characteristic of a market where transactions do not markedly move prices. Liquid markets usually have low *bid–ask spreads* and high volume of transactions.

Limit order Order to buy a specified quantity up to a maximum price, or sell subject to a minimum price.

Market impact cost Cost of a trade resulting from an order moving the price against a trader. Such costs are smaller in a market with high *liquidity*.

Market-maker *Dealer* obliged to quote buy and sell prices in return for certain privileges within a market.

Microstructure (market) Literature analysing specific institutions, exchanges and trading rules, especially their role in price discovery.

NASDAQ *OTC* market for US equities specializing in high-tech stocks.

Network effect, network externalities Tendency for liquid markets to attract further liquidity as market participants want to trade where others are already actively trading. Sometimes referred to as 'demand side *economies of scale*'.

NYSE New York Stock Exchange.

Opacity Opposite of *transparency*.

Order book A market where prices are determined by an order execution algorithm from participants sending firm buy and sell orders.

Order routing Delivery of messages from end users to the *execution* system.

OTC Over the counter. Bilateral transactions not conducted on a formal exchange.

Protocol A set of rules governing trading, e.g. the types of orders allowed, minimum tick size, rules to halt trading, special rules for openings and closings.

Public good Commodity or service available to everyone in an area, regardless of their willingness to pay, and where one person's consumption does not diminish that of others. Classic examples include defence and street lighting.

Quality (market) Broadly, market performance across a range of attributes such as *liquidity*, volatility and *resilience*.

Resilience Ability of a market to function in an efficient, *liquid* and orderly manner at times of great price uncertainty and market stress.

SEC Securities and Exchange Commission. A major regulator of financial markets in the USA.

Settlement Completion of a transaction by exchange of instrument and funds.

STP Straight-through processing: the capture of trade details directly from frontoffice systems to backoffice. Completes automated processing of *clearing* and *settlement* instructions without the need for rekeying or reformatting data.

Tipping Tendency for a system provider that has achieved a large market share to move quickly to a (near) monopoly.

Transparency Ability of market participants to observe trade information in a timely fashion.

Upstairs trading Special arrangement for handling (usually) *block trades*, possibly involving less transparency.

Notes

1 There is now a range of work concerning aspects of electronic trading, as references throughout this paper attest. There is also a considerable amount of work on the effect of technology on market structures and on electronic finance more generally – e.g. Allen, McAndrews and Strahan (2002), Bar (2001), Mishkin and Strahan (1999), Claessens, Glaessner and Klingebiel (2000) and Sato and Hawkins (2001).

2 Market microstructure literature studies the processes/outcomes of exchanging assets under explicit trading rules – O'Hara (1995) provides a theoretical review; see also the recent survey by Madhavan (2000b).

3 Financial systems are integral to the funding of capital accumulation, one of the fundamental drivers of economic growth. Greater efficiency in infrastructure mechanisms such as trading systems can reduce frictions in financial systems, helping to bring savers and investors together more efficiently, thus aiding growth. Links between the functioning of financial systems and growth are explored in Levine (1997) and more recently, for example, in Leahy *et al*. (2001) with empirical analysis.

4 The last three out of the four stages characterized are similar to that of the traditional 'structure/conduct/performance' framework from industrial organization literature – see, for example, the textbook by Carlton and Perloff (1999). The process indicated in the flowchart and Box A may well be iterative, as welfare improvements stimulate further innovation, and so on.

5 While these particular examples have been chosen to illustrate patterns of developments, to a lesser or greater extent all financial markets have been influenced by electronic trading developments. For example, Tsang (1999) reviews automation in futures trading, where electronic trading has been well-established for some time. Banks (2001, Chapter 4) describes electronic trading in a range of markets, with particular focus on the role of the Internet.

6 The regulatory trigger was the SEC's order-handling rules, implemented in January 1997. McAndrews and Stefanadis (2000) set out the regulatory and technological developments that have affected the ECNs. Separately, a range of other electronic trading systems offer a considerable variety of market architecture. For a further discussion of the US experience, see Davis and Steil (2001, Chapter 8) and Domowitz and Steil (2001b).

7 See, for example, Cavaglia, Brightman and Aked (2000).

8 See, for example, *The Economist* (2001a, 2001b). A list of automated exchange mergers and alliances over 1997–2000 is presented in Davis and Steil (2001, Table 7.2). See also the discussion of stock exchange alliances in Gaa *et al*. (2001, p. 54).

9 In a typical telephone dealer market (e.g. that for UK government securities), customers telephone orders to dealers who either are obliged to quote prices or do so on a 'best efforts' basis. The dealers manage their inventory in an inter-dealer market, sometimes intermediated by inter-dealer brokers and sometimes dealing directly with other dealers.

10 The comparison of the UK equity and gilt markets found in Annex A of DMO (2000) suggests that the average gilt market trade was approximately thirty times that in the equity market. Differences between the characteristics of government bond and equity markets are detailed in Gravelle (2002).

11 Claessens, Glaessner and Klingebiel (2001) give examples of platforms for trading in bonds from a number of emerging economies, including Asian bonds, Latin American bonds and Brady bonds. Bank of Japan (2001) notes that since mid-2000 a number of bond trading systems have been established in Japan, but trading volumes are still low.

12 Fixed income markets are much less studied than are equity markets: Goodhart and O'Hara (1997) noted that 'the absence of much market microstructure analysis in (government) bond markets is particularly surprising', a comment echoed by Lyons (1998) and Gravelle (2002). One recent study is Study Group on Fixed Income Markets (2001).

13 Across the whole trade processing infrastructure, there are wider consolidation issues. Notably, there is considerable scope for consolidation in clearing and settlement procedures, progress on which may be related to exchange consolidation – see, e.g., Gaa *et al*. (2001).

14 Current pressures affecting consolidation in the financial sector as a whole, including its patterns, causes and policy implications, are examined in the report by the Group of Ten (2001), and in emerging economies by Hawkins and Mihaljek (2001). The effect of technology and the Internet on fragmentation and consolidation in US equity markets is explored in Madhavan (2000a).

15 Longer opening hours may also seem to be cheaper and more feasible – but in practice longer (even 24-hour) trading sessions do not seem valued in wholesale markets. Reasons may include a preference for a distinct end-of-day closing for risk management procedures and, more fundamentally, lack of demand from customers and the problems of thin markets.

16 Moreover, moving to different or additional systems is far from friction-free. For example, at the most practical level it may involve extra screens, linkages and staff training. While it is likely that such costs will progressively lessen, the point at which they can be virtually ignored seems distant.

17 This section focuses on transparency in terms of the ability of market participants to observe information about the trading process. As discussed in O'Hara (1995), there are multiple dimensions to transparency. Consideration of it can be split into *pre-trade* information on order sizes and quotes, and *post-trade* information on prices and quantities of executed trades. Other considerations include the timeliness of the information made available, which (a subset of) participants can observe certain aspects, and pre- and post-trade anonymity (whether identities are revealed).

18 Madhavan (2000b) surveys results regarding transparency from theoretical, empirical and experimental literature. Much of the work uses underlying models based on asymmetric information – these consist of two classes of market participants, informed traders with private information on future asset values and uninformed (liquidity motivated) traders, and explore how these groups trade under different conditions. Such models are mostly applicable to equity markets in which private information on assets plays an important role. There is also a range of models based around inventory

adjustment, consisting of dealers who attempt to restore their inventories to some desired level by adjusting their quotes and trading behaviour. As discussed in Gravelle (2002), these also fit closer with the structures typically seen in fixed income and foreign exchange markets.

19 Such systems are a response to the above point, that in a transparent environment, wholesale traders may well disguise these orders in some way to avoid giving away information on their strategy which may lead to the market moving against them. The pre-trade opaque class of systems in contrast actually allow traders to input their true order preferences to the system with complete accuracy since the information is only 'seen' by the computer system. These systems aim to meet a demand for trading without losing informational advantage.

20 Considering transparency more widely, other forms of information dissemination also become more feasible as technology is integrated. For example, information on the characteristics of numerous securities can be made available (as seen in the systems implemented in several government bond markets) while transparency of processes, such as order routing information, can also be offered. However, questions have been raised (see D'Avolio, Gildor and Schleifer, 2001 and Ferguson, 2001) about the extent to which the increased quantity of information enabled by technology is accompanied by an increase in its quality.

21 Trading cost studies typically measure costs conditional on execution. There are also costs arising from non-execution, which particularly affect the placement of limit orders. There are also opportunity costs associated with any trading decision.

22 See, for example, the articles listed by Domowitz and Steil (2001a). Domowitz (2002) claims 'all things equal, average trading costs are lower by thirty-three to forty-six basis points in markets which are largely automated'. Jiang, Tang and Law (2002) show that the introduction of electronic trading in the Hong Kong futures market lowered bid–ask spreads, even after allowing for changes in price volatility and trading volume.

23 In electronic systems, price is usually the first criterion, priority being accorded to the highest bids and lowest offers. The most common secondary criterion is time, with orders transacting on a 'first come, first served' basis. Other systems spread a new matching bid *pro rata* over all the offers at that price. In some systems, market orders may be given priority over limit orders, large orders over small orders, openly disclosed orders over anonymous orders, or those from designated market-makers over those of other traders. Another possibility would be for a system to trade off automatically price *versus* credit risk in accepting bids. Domowitz (1993) provides a taxonomy of these algorithms and describes those used by systems in the early 1990s.

24 For examples of market comments about electronic trading possibly degrading the quality of prices, see Morris (2001) and McNee (2000).

25 There are other instances – not necessarily associated with electronic trading – where prices in one market are largely determined by prices in another, possibly more liquid, market. Examples include off-the-run bonds, whose price movements largely reflect those of on-the-run bonds. In Japan, yields in the government bond market are largely driven by developments in the more liquid futures market. Between the polar cases of pure price making and price taking there are also systems that set the price within limits of prices from a central market.

26 Conrad, Johnson and Wahil's (2001) study of the US equity market finds that for crossing networks average implicit cost is 0.23 per cent and explicit cost 0.07 per cent compared with 0.54 per cent and 0.27 per cent respectively for full service brokers.

27 Characteristics of liquidity in markets are discussed in the policy context in BIS (2001) and from a theoretical perspective in O'Hara (1995). There is discussion of the impact of the Internet on equity market liquidity in Madhavan (2000a). Some evidence on electronic trading and market resilience is discussed in CGFS (2001a,b).

28 Analogies can be made with other areas of e-finance, such as the difficulties and/or high costs of Internet-only banks attracting funds; see de Young (2001).

29 Clearly the enhanced possibilities for market access offered by electronic trading themselves can garner liquidity. Electronic trading's other effects on liquidity can come via influence on fragmentation and consolidation and, connected to this, its influence in allowing trades to be executed away from main markets, for example by internalization of order flow. Lower trading costs in themselves also attract more trading.

30 For example, US SEC (1998); proposals by the Canadian Securities Administrators (2000) concerning alternative trading systems – discussed in Boisvert and Gaa (2001); discussions in the United Kingdom – see Financial Services Authority (2000) and Wisbey (2000). This section highlights some key, underlying issues. Details about the debates and the complexity of considerations can be found in the documents referenced.

31 Three main categories of network bias are described by Bar (2001), as follows. Information asymmetries are where the market is structured so some players have better/earlier market-relevant information. Matching asymmetries are where market clearing algorithms are programmed to favour some participants. Access asymmetries are where different players have differential access to the telecommunications infrastructure.

32 Several official speeches elaborate on these points, for example Crockett (2001) and Greenspan (2000).

References

Allen, F., McAndrews, J. and Strahan, P. (2002) 'E-finance: an introduction', *Journal of Financial Services Research*, 22 (1): 5–27.

Allen, H. and Hawkins, J. (2002) 'Electronic trading in wholesale financial markets: its wider impact and policy issues', *Bank of England Quarterly Bulletin* 42(1): 50–8.

Bank for International Settlements (2001) *Market Liquidity: Proceedings of a Workshop held at the BIS*, BIS Papers 2, April. Online <http://www.bis.org/publ/bispap02.htm> (accessed 6 November 2001).

Bank of England (2000) *Oversight of Payment Systems*, November. Online <http://www.bankofengland.co.uk/fsr/ops.pdf> (accessed 6 November 2001).

Bank of Japan (2001) 'Increasing use of electronic trading systems and its implications on Japanese financial markets', *Market Review*, 2001-E-2, July, Financial Markets Department, Bank of Japan. Originally published in Japanese as 2001-J-1 in January 2001. Online <http://www.boj.or.jp/en/ronbun/ronbun_f.htm> (accessed 6 November 2001).

Banks, E. (2001) *E-Finance – the Electronic Revolution*. John Wiley.

Bar, F. (2001) 'The construction of marketplace architecture', in S. Cohen and J. Zysman (eds) *Tracking a Transformation: E-commerce and the Terms of Competition in Industries*. Berkeley.

Boisvert, S. and Gaa, C. (2001) 'Innovation and competition in Canadian equity markets', *Bank of Canada Review*, Summer: 15–30. Online <www.bankofcanada.ca/en/res/r01-3-eb.htm> (accessed 6 November 2001).

Bond Market Association (2001) *eCommerce in the Fixed Income Markets: the 2001 Review of Electronic Transaction Systems*, December. Online <http://www.bondmarkets .com/research/ecommerce/ETSRprt1201.pdf> (accessed 2 August 2002).

Breedon, F. and Holland, A. (1998) 'Electronic versus open outcry markets: the case of the Bund futures contract', Bank of England Working Paper no. 76, February. Online <http://www.bankofengland.co.uk/workingpapers/wp76.pdf> (accessed 5 November 2001).

Canadian Securities Administrators (2000) 'Alternative trading system proposal', *OSC Bulletin* 23, July.

Carlton, D. and Perloff, J. (1999) *Modern Industrial Organisation*. Longman.

Cavaglia, S., Brightman, C. and Aked, M. (2000), 'The increasing importance of industry factors', *Financial Analysts Journal* 56: 41–54.

Chaboud, A. and Weinberg, S. (2002) 'Foreign exchange markets in the 1990s: intraday market volatility and the growth of electronic trading', in *Market Functioning and Central Bank Policy*, BIS Papers 12.

Claessens, S., Glaessner, T. and Klingebiel, D. (2000) 'Electronic finance: reshaping financial landscapes around the world', Financial Sector Discussion Paper 4, World Bank, September.

Claessens, S., Glaessner, T. and Klingebiel, D. (2001a) 'E-finance in emerging markets: is leapfrogging possible?', Financial Sector Discussion Paper 7, World Bank, June.

Claessens, S., Glaessner, T. and Klingebiel, D. (2001b) 'Structural aspects of market liquidity from a financial stability perspective', Discussion note no. 1, Basel, June. Online <http://www.bis.org/publ/cgfs_note01.htm> (accessed 5 November 2001).

Committee on the Global Financial System (CGFS) (2001), *The Implications of Electronic Trading in Financial Markets*, no. 16, Basel, January. Online <http://www.bis.org/publ/cgfs16.htm> (accessed 5 November 2001).

Conrad, J., Johnson, K. and Wahil, S. (2001) 'Institutional trading and soft dollars', *Journal of Finance* LVI: 397–416.

Crockett, A. (2001) 'Market discipline and financial stability', speech at the Banks and Systemic Risk Conference, Bank of England, London, 23–25 May. Online <http://www.bis.org/speeches/sp010523.htm> (accessed 5 November 2001).

Davis, E. and Steil, B. (2001) *Institutional Investors*. London: MIT Press.

D'Avolio, G., Gildor, E. and Schleifer, A. (2001) 'Technology, information production and market efficiency', paper presented at Federal Reserve Bank of Kansas City symposium on *Economic Policy for the Information Economy*, August–September. Online <http://www.kc.frb.org./Publicat/sympos/2001/papers/S02shle.pdf> (accessed 27 August 2002).

Debt Management Office (DMO) (2000) *The Secondary Market for Gilts: a Consultation Paper*, 24 January. Online <http://www.dmo.gov.uk/publication/g2gil.htm> (accessed 6 November 2001).

Degryse, H. and van Achter, M. (2001) 'Alternative trading systems and liquidity', Chapter 10 of this volume.

De Young, R. (2001) 'The financial progress of pure-play internet banks', BIS Papers 7, pp. 80–6, November. Online <http://www.bis.org/publ/bispap7.htm>

Domowitz, I. (1993) 'A taxonomy of automated trade execution systems', *Journal of International Money and Finance* 12: 607–31.

Domowitz, I. (2002) 'Liquidity, transactions costs and reintermediation in electronic markets', *Journal of Financial Services Research* 22(1): 141–57.

Domowitz, I., Glen, J. and Madhavan, A. (2001) 'Liquidity, volatility and equity trading costs across countries and over time', *International Finance* 4(2): 221–55. Online <http://www.smeal.psu.edu/faculty/ihd1/intcost1.pdf> (accessed 6 November 2001).

Domowitz, I. and Steil, B. (2001a) ‘Automation, trading costs, and the structure of the securities trading industry’, Chapter 7 of E. Davis and B. Steil, *Institutional Investors*. London: MIT Press. (An earlier version appeared in R. Litan and A. Santomero (eds), *Brooking-Wharton Papers on Financial Service*, 1999, pp. 33–81.)

Domowitz, I. and Steil, B. (2001b) ‘Innovation in equity trading systems: the impact on transactions costs and cost of capital’, in R. Nelson, D. Victor and B. Steil (eds), *Technological Innovation and Economic Performance*, Princeton University Press. Online <http://www.smeal.psu.edu/faculty/ihd1/innovation1.pdf> (accessed 5 November 2001).

Economides, N. (1996) ‘The economics of networks’, *International Journal of Industrial Organisation* 16: 673–99. Online <http://www.raven.stern.nyu.edu/networks/top.html> (accessed 5 November 2001).

Economides, N. (2001) ‘The impact of the internet on financial markets’, *Journal of Financial Transformation* 1(1). Online <http://www.stern.nyu/networks/The_impact_of_the_internet_on_financial_markets.pdf> (accessed 6 November 2001).

Economist, The (2000) ‘E-bonds, licensed to kill’, 15 January.

Economist, The (2001a) ‘Rise and fall: a survey of global equity markets’, 5 May.

Economist, The (2001b) ‘The hunt for liquidity’, 28 July.

Fan, M., Stallaert, J. and Whinston, A. (2000) ‘Designing electronic market institutions for bond trading’, in E. Melnick, P. Nayyer, M. Pinedo and S. Seshadri (eds), *Creating Value in Financial Services*. Boston: Klewer Academic Publishers.

Ferguson, R. (2001) ‘Remarks’ at Federal Reserve Bank of Kansas City symposium on *Economic Policy for the Information Economy*, August–September. Online <http://www.kc.frb.org/publicat/sympos/2001/sym01prg.htm> (accessed 5 November 2001).

Financial Services Authority (2000) ‘The regulation of market infrastructure’, January. Online <http://www.fsa.gov.uk/pubs/discussion/d02.pdf> (accessed 5 November 2001).

Frino, A., McInish, T. and Toner, M. (1998) ‘The liquidity of automated exchanges: new evidence from German Bund futures’, *Journal of International Financial Markets, Institutions and Money* 8: 225–41.

Gaa, C., Ogrodnick, R., Thurlow, P. and Lumpkin, S. (2001) ‘Future prospects for national financial markets and trading centres’, *Financial Market Trends*, OECD no. 78, March.

Ganley, J., Holland, A., Saporta, V. and Vila, A. (1998) ‘Transparency and the design of securities markets’, *Financial Stability Review*, Bank of England, 4: 8–17.

Goodhart, C. and O’Hara, M. (1997) ‘High frequency data in financial markets: issues and applications’, *Journal of Empirical Finance*, 4: 73–114.

Gravelle T. (2002) ‘The microstructure of multiple-dealer equity and government securities markets: how they differ’, Bank of Canada Working Paper 2002–9. Online <http.//www.bankofcanada.ca/en/res/wp02-9.htm> (accessed 2 August 2002).

Green, C., Maggioni, P. and Murinde, V. (2000) ‘Regulatory lessons for emerging stock markets from a century of evidence on transactions costs and share price volatility in the London Stock Exchange’, *Journal of Banking and Finance* 24: 577–602.

Greenspan, A. (2000) ‘Electronic finance’, remarks to the 9th Annual Financial Markets Conference of the Federal Reserve Bank of Atlanta, 16 October. Online <http://www.federalreserve.gov/boarddocs/speeches/2000/20001016.htm> (accessed 5 November 2001).

Group of Ten (2001) *Report on Consolidation in the Financial Sector*, Basel, January. Online http://www.bis.org/publ/gten05.htm (accessed 5 November 2001).

Hawkins, J. and Mihaljek, D. (2001) 'The banking industry in the emerging market economies: competition, consolidation and systemic stability: an overview', *BIS Papers* no. 4, September. Online http:/www.bis.org/publ/bispap04.htm (accessed 5 November 2001).

Hendershott, T. and Mendelson, H. (2000) 'Crossing networks and dealer markets: competition and performance', *Journal of Finance* LX: 2071–115.

Jiang, G., Tang, N. and Law, E. (2002) 'Electronic trading in Hong Kong and its impact on market functioning', in *Market Functioning and Central Bank Policy*, BIS Papers 12.

Leahy, M., Schich, S., Wehinger, G., Pelgrin, F. and Thorgeirsson, T. (2001) 'Contributions of financial systems to growth in OECD countries', Economics Department Working Paper no. 280, OECD. Online <http://www.oecd.org/pdf/M00002000/M00002479.pdf> (accessed 12 November 2001).

Lin, L., Geng, X. and Whinston, A. (2001) 'A new perspective on finance and challenges for financial institutions in the internet era', BIS Papers 7, pp. 13–25, November. Online <http://www.bis.org/publ/bispap7.htm>

Lee, R. (1998) *What is an Exchange*?. Oxford University Press.

Levine, R. (1997) 'Financial development and economic growth: views and agenda', *Journal of Economic Literature* 35: 688–726.

Long, S. (2000) 'Online finance: the virtual threat', *The Economist*, 20 May.

Lyons, R. (1998) 'Introduction to the international market microstructure issue', *Journal of International Financial Markets, Institutions and Money* 8: 219–23.

Madhavan, A. (1996) 'Security prices and market transparency', *Journal of Financial Intermediation* 5: 255–83.

Madhavan, A. (2000a) 'In search of liquidity in the internet era', presented to the 9th Annual Financial Markets Conference of the Federal Reserve Bank of Atlanta, October. Online <http://www.frbatlanta.org/news/conferen/efinance00/madavan_final.pdf> (accessed 5 November 2001).

Madhavan, A. (2000b) 'Market microstructure: a survey', *Journal of Financial Markets* 3: 205–58.

McAndrews, J. and Stefanadis, C. (2000) 'The emergence of electronic communications networks in the US equity markets', *Current Issues in Economics and Finance*, Federal Reserve Bank of New York, 6, 12: 1–6. Online <http://www.newyorkfed.org/rmaghome/curr_iss/ci6-12.html> (accessed 5 November 2001).

McNee, A. (2000) 'Looking for a central system', *Risk*, February: 55–6.

Mishkin, F. and Strahan, P. (1999) 'What will technology do to financial structure?', in R. Litan and A. Santomero (eds), *Brookings-Wharton Papers on Financial Services*, Washington DC: Brookings Institution Press.

Morris, J. (2001) 'Forex transformed by mergers', *Euromoney*, May: 61–8.

Moszkowski, G., Bedell, B., Vetto, M. and Clark, C. (2001), *Trading up 2.0: The New World of Electronic Trading.* New York: Salomon Smith Barney, 22 March.

O'Hara, M. (1995) *Market Microstructure Theory*. Oxford: Blackwell.

Picot, A., Bortenlaenger, C. and Roehrl, H. (1995) 'The automation of capital markets', *Journal of Computer-Mediated Communication* 1(3). Online <http://www.ascusc.org/jcmc/vol1/issue3/picot.html> (accessed 5 November 2001).

Pirrong, S. (1996) 'Market liquidity and depth on computerised and open-outcry trading systems: a comparison of DTB and LIFFE bund contracts', *Journal of Futures Markets* 16: 519–43.

Sato, S. and Hawkins, J (2001) 'Electronic finance: an overview of the issues', BIS Papers, no. 7: 1–12.

Sato, S., Hawkins, J. and Berentsen, A. (2001) 'E-finance: recent developments and policy implications', in S. Cohen and J. Zysman (eds), *Tracking a Transformation: E-commerce and the Terms of Competition in Industries*. Berkeley.

Schwartz, R. (ed.) (2001) *The Electronic Call Auction: Market Mechanism and Trading: Building a Better Stock Market*. Kluwer Academic Publishers.

Shapiro, C. and Varian, H. (1999) *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press.

Soejima, Y. (2001) 'Order-matching system and order strategy in JGB futures market', Bank of Japan Financial Markets Working Paper 01-J-1 (in Japanese, summarized in Box E of CGFS, 2001).

Steil, B. (2001) 'Borderless trading and developing securities markets', in R. Litan, P. Masson and M. Pomerleano (eds), *Open Doors: Foreign Participation in Financial Systems in Developing Countries*. Washington: Brookings Institution Press.

Study Group on Fixed Income Markets (2001) 'The changing shape of fixed income markets', BIS Working Paper 104, September. Online <http://www.bis.org/publ/work104.htm> (accessed 5 November 2001).

Tsang, R. (1999) 'Open outcry and electronic trading in futures markets', *Bank of Canada Review*: Spring, 21–39. Online <http://www.bankofcanada.ca/en/res/r992b-ea.htm> (accessed 6 November 2001).

Turner, P. (2001) 'E-finance and financial stability', in R. Litan, P. Masson and M. Pomerleano (eds) *Open Doors: Foreign Participation in Financial Systems in Developing Countries*. Washington: Brookings Institution Press.

US Securities and Exchange Commission (1998) 'Regulation of exchanges and alternative trading systems', Release 34-40760, December. Online <http://www.sec.gov/rules/final/34-40760.txt> (accessed 6 November 2001).

Vila, A. and Sandman, G. (1995) 'Floor trading *versus* electronic screen trading: an empirical analysis of market liquidity and information transmission in the Nikkei stock index futures market', London School of Economics Financial Markets Group Discussion Paper 218, October.

Wisbey, G. (2000) 'The challenges of technology – regulation of electronic financial markets', speech to *The Challenges Facing Financial Regulation*, University of Cambridge, Centre for Corporate and Commercial Law, 6–7 July. Online <www.fsa.gov.uk/pubs/speeches/sp53.html> (accessed 12 November 2001).

Acknowledgements

The authors particularly thank Simon Buckle, Charles Gaa, Gabriele Galati, Joe Ganley, Giuseppe Grande, Toni Gravelle, Serge Jeanneau, Jamie McAndrews, Tom Minic, Inke Nyborg, Sarah Parkinson, David Rule, Clifford Smout, John Trundle, Philip Turner and participants at the SUERF colloquium for helpful comments on this paper. In addition, the authors all benefited from discussions in 2000 while participating in the Electronic Trading Working Group of the G10 Committee on the Global Financial System. The opinions expressed are those of the authors and not necessarily shared by the Bank of England, the Bank for International Settlements or its other member central banks. A revised version of this paper has been published in the Spring 2002 issue of Bank of England's *Quarterly Bulletin*.

Part IV

Contributions on technology and payments

13 Monetary policy in a cashless society*

Cláudia Costa Storti and Paul De Grauwe

1 Introduction

Innovations in information technologies have improved the prospects for the advent of a cashless society. Although it remains unclear whether the new information technologies will drive out cash completely from the payments system in the foreseeable future, the prospects for such an outcome cannot be excluded either.[1]

In this paper we study the implications of a cashless society for the conduct of monetary policies. We define a cashless society to have the following characteristics. One is that there are no notes and coins in circulation issued by a central bank. All the money used is issued by private financial institutions (banks and possibly other firms). It is conceivable that the central bank continues to operate like other banks issuing its own deposits that could be used as money in the same way as other bank deposits are. However, in that case the central bank has no monopoly in the issue of money. We will continue to assume that in a cashless society the unit of account (e.g. one dollar, one euro) remains a national affair and is provided by the state. Thus banks in the same country issue deposits, or other means of payments, in the same national unit of account.[2] As we will argue, this will require the state to exert some prudential supervision over the institutions that issue money.

We will analyse monetary policies in a cashless society in two steps. In a first step we study the following problem: is it possible that the private issue of money will lead to an unstable price level? In other words will there be a mechanism that ties down the price level and prevents systematic inflation? This problem has been called the *price indeterminacy problem*. We analyse it in Section 2.

In the second step we study what the role will be of the central bank (if any) in a cashless society. In particular, we will analyse whether and how the central bank can be transformed so that it can maintain price stability in a cashless society.

2 The problem of price indeterminacy

The problem of price indeterminacy has been analysed extensively by economists. One of the earliest writers on this issue was Irvin Fisher (1896). Later Patinkin (1965) and Fama (1980) added important insights.

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

The classical view of the indeterminacy of the price level can be formulated in a simplified way as follows.[3] We start from the demand and the supply of money.

The demand for money can be specified as

$$P\,L(Y,r) \tag{1}$$

where P is the price level, L is the real demand for money, which itself is a function of output, Y, and the nominal interest rate, r. Note that we assume that the demand for money is homogeneous of degree one. This also implies that agents are free of money illusion.

The supply of money is M. The mechanism by which the supply of money is generated will turn out to be of crucial importance.

Money market equilibrium is achieved when the demand equals the supply, i.e.

$$M = P\,L(Y,r) \tag{2}$$

or

$$\frac{M}{P} = L(Y,r) \tag{3}$$

This very simple equation describing monetary equilibrium can be used to illustrate the potential for price indeterminacy. The right-hand side of the equation represents the *real* demand for money. It is the result of what happens in the 'real' economy, i.e. transactions in the markets for goods and services lead to a demand for money balances. Note the importance of the absence of money illusion: agents do not care whether additional money desires are realized by increases in the money stock or declines in the price level.

The left-hand side of (3) is the real supply of money. We can immediately establish the following basic result: for any given real demand for money there are infinitely many combinations of money stock and price levels that will do the job of bringing about money market equilibrium.

The root of the indeterminacy problem can also be formulated more generally.[4] Private agents that do not suffer money illusion, care only about relative prices. They do not care about nominal variables such as the price level, or the nominal money stock. If there is nobody else in the system who cares about nominal variables, the latter will not be determinate. They can take any value. It is important, therefore, that some institution exists that will take the responsibility to fix nominal variables. In a cash society this responsibility is assumed by the central bank. We analyse how modern central banks do this in the next section.

2.1 The money supply in a cash society

In present-day economies money consists of both currency and bank deposits. There is a consensus today that central banks have the capacity to control the price level in two ways.[5] One approach consists in controlling the money supply. This approach has been advocated by the monetarist school and has led a number of central banks to implement money supply targeting procedures. The best known example was the German Bundesbank.[6] This approach has encountered a number of practical problems, however, that have led some central banks to drop it (e.g. the US Federal Reserve, the Bank of England). Today few central banks rely on this. The only noteworthy exception is the ECB (although it does not like to use the term 'money supply targeting').

A second approach has found increasing reliance among central bankers. This consists in following a Taylor-type rule whereby the central bank adjusts the short-term interest rate in response to movements in expected inflation and the state of economic activity (most often measured by the output gap). This rule typically implies that when expected inflation increases, the central bank raises the nominal interest rate sufficiently to generate an increase in the real interest rate. This in turn helps to reduce inflationary pressures. There is now increasing evidence that this is what most central banks do in one way or the other (see Taylor, 1993 and Clarida, Gali and Gertler, 1997).[7] The empirical evidence also indicates that most central banks following this procedure have been quite successful in keeping inflation low.

From the preceding discussion one can conclude the following. Central banks in a cash society have the tools to solve the price indeterminacy problem. They can either control the money stock, which will then tie down the price level, or they can try to control the rate of change of the price level, thereby avoiding the indeterminacy problem. In a cash society there is a 'guardian' who cares about nominal variables and who has the tools to control these nominal variables.

2.2 The money supply in a pure cashless economy

We now analyse the issue of whether in a cashless society there exists a similar mechanism that makes it possible to fix the price level. As will be remembered, we define a cashless society to be a regime in which currency issued by the central bank has ceased to exist. All the money is private money issued by banks in the form of deposits, or some fancier e-money issued by institutions that are not necessarily banks. We call the cashless regime *pure* because we will assume here that the central bank does not impose reserve requirements. In such a pure cashless society, banks and other institutions issuing money do not hold reserves at the central bank. (We will come back to this issue and ask the question what difference it makes when the central bank would impose minimum reserve requirements on issuers.)

The problem of the indeterminacy of the price level was analysed by classical economists. They analysed the issue of whether a monetary regime based on private bank money (deposits) would be capable of stabilizing the price level. The consensus view was that this was not the case.[8]

Let us analyse the situation of a country where the only form of money is demand deposits issued by private banks. Currency has disappeared from circulation and there is no central bank issuing money. In order to analyse such a monetary regime it is useful to start from the banks' balance sheet:

$$A_t = M_t \tag{4}$$

where A_t is the value of the portfolio of financial assets held by the banks in period t and M_t is the value of the deposits issued by the banks in period t. It is assumed that deposits are the only means of payments in the economy. This balance sheet makes very clear that the process by which money is created is linked to the process by which credit is created. We can now develop the fundamental theorem of indeterminacy of the price level in a cashless economy. First divide (4) by the price level P_t and combine with (3):

$$\frac{A_t}{P_t} = \frac{M_t}{P_t} = L(Y_t, r_t) \tag{5}$$

It follows that the price indeterminacy can be eliminated if A_t, the portfolio of assets, can be pinned down. But can it? The answer is that in a cashless economy it cannot. In order to see this, let us take the easy case first, i.e. the banks' portfolio of assets consists of shares:

$$A_t = S_t Q_t \tag{6}$$

where Q_t is the number of shares held in the banks' portfolio at time t and S_t is the market price of shares at time t. We now can use the standard theory about how the price of shares is determined. This theory says that the price of a share today is equal to the present value of the future dividend stream, i.e.

$$S_t = \sum_{i=1}^{\infty} \left(\frac{1}{1+r}\right)^i E_t d_{t+i} \tag{7}$$

where E_t is the expected value operator and d_{t+i} is the dividend in period $t+i$.

These future dividends are influenced by many variables. One prominent variable is the money stock. The higher the money stock the higher will be the value of these future dividends. The reason is that with a higher money stock nominal values (prices) increase, including the nominal value of dividends. Thus we can write:

$$d_t = \mathrm{f}(M_t) \tag{8}$$

i.e. the nominal value of dividends is a (positive) function of the money stock. When the money stock increases, the *nominal* value of dividends also increases. Substituting (8) into (7) and (7) into (6) yields:

$$A_t = \left[\sum_{i=1}^{\infty} \left(\frac{1}{1+r}\right)^i E_t \, \mathrm{f}(M_{t+i}) \right] Q_t \tag{9}$$

We conclude that the value of the banks' assets depends on the expected future deposits (money stock). Since the bank balance sheet constraint implies $A_t = M_t$, we have that

$$M_t = \left[\sum_{i=1}^{\infty} \left(\frac{1}{1+r} \right)^i E_t \, \mathrm{f}(M_{t+i}) \right] Q_t \tag{10}$$

It follows that the stock of deposits (money) today depends on the expected future stock of deposits (money). Any expectation about the future level of deposits (money) is validated.

The mechanism underlying this self-fulfilling expectations process can be described as follows. If agents expect higher dividends in the future, this raises share prices today and thus the value of the banks' assets today. This in turn increases the value of the banks' deposits. Since deposits perform the role of money, the higher money stock validates a higher future dividend stream. The money stock can take on any value depending on the agents' expectations about future dividend streams. Nothing will tie down the money stock in this system. Using equation (5) we can also conclude that the price level is indeterminate.

This feature of the banking system creates a potential for inflation both in share prices and in the general price level. Because of their self-fulfilling character, expectations can also lead to speculative bubbles in the stock prices. This is also the reason why in many countries banks have been forbidden to hold shares in their portfolio.

Restricting the types of assets that banks can hold in their portfolio to fixed interest securities (e.g. bonds, mortgages) eliminates an important source of price indeterminacy. Assume that banks would only hold bonds in their portfolio. We now redefine S_t in equation (6) to be the price of bonds and Q_t the quantity of bonds held by the banks. The price of bonds to be repaid in period T can be written as

$$S_t = \sum_{i=1}^{T} \left(\frac{c}{1+r} \right)^i + \frac{K}{(1+r)^T} \tag{11}$$

where c is the fixed coupon of the bond and K is the face value of the bond. Since c and K are fixed numbers the only uncertainty arises from changes in the interest rate used in computing the present value and the default risk. In contrast to the pricing of stocks, there is no expectations mechanism that can lead to self-fulfilling behaviour of the price of bonds. Instead there is a strong mean reversion process underlying the bond prices, i.e. as we come closer to time T when the bond matures, the bond price must return to its face value. Thus in this case, S_t is tied down, and so is A_t. Does this mean that the price indeterminacy problem can be eliminated by imposing a condition that banks can only hold fixed interest securities in their portfolio? The answer is negative. To see why we also have to analyse the link between the money stock and Q_t (the number of securities in the banks' portfolio).

There are other mechanisms linking the money stock to the banks' assets. One such mechanism goes from money to Q_t, the amount of bank assets (loans). It can be described as follows. When the money stock increases this stimulates economic activity. As a result firms will want to borrow more and issue more bonds (or other fixed return securities). There will also be a higher (transactions) demand for money. It follows that the balance sheet of the banks will increase, i.e. banks will accumulate more bonds and issue more deposits. Q_t increases.

A second mechanism recognizes that bank loans are often extended based on adequate collateral. The example is mortgage loans. This provides a link between the money stock and the value of the collateral, in this case the value of real estate. When the money stock increases this will lead to higher asset prices, including higher real estate prices. These higher real estate prices then increase the value of the collateral, which in turn increases the value of new loans. Again a higher money stock leads to an increase in the value of the banks' assets.

We conclude that although the prices of fixed interest securities (including mortgage loans) are tied down there exist other mechanisms that lead to an indeterminacy of the value of the banks' assets and thus of the price level. This mechanism is especially relevant with respect to mortgage loans. Time and again, banks have been willing to increase their portfolio of mortgage loans in response to higher real estate prices. These higher real estate prices, however, have often been caused by excessive money creation. This has often led to bank crises and collapses. In the absence of a central bank this feature of private banking could go on unchecked.

2.3 Conclusion

We have shown that in a monetary system where all money is provided by private institutions there is a potential for price indeterminacy and inflation. In other words a monetary system without an outside 'guardian' who cares about nominal variables will find it difficult to avoid the problem of price indeterminacy. The question, therefore, becomes the following. Can the central bank in a future cashless society (assuming that central banks will continue to exist) assume its present role of stabilizing the price level? And if so, how?

We analyse these questions in the next sections. In Section 3 we assume that the central bank does not impose legal reserve requirements. As will be remembered, we call such a system a *pure* cashless society. In Section 4, we allow for the possibility that the central bank would impose legal reserve requirements on private banks issuing deposits, and on all other institutions that provide new forms of money (e-money, Internet money, etc.). Many central banks do this today, although for reasons that are not always related to monetary control.[9] It is conceivable, even probable, that central banks will want to continue to use reserve requirements.

A final point should be stressed here. We mentioned earlier that in a cashless society the central bank or another official institution is likely to exert powers of supervision on those institutions that issue private money. In Section 4 we will also study the issue of how this supervisory function can be strengthened so as to maintain monetary and price stability in a cashless society.

3 The central bank in a pure cashless society

In the previous section we came to the conclusion that a private system of money supply faces the problem that the price level cannot easily be controlled. This raises the issue of how the price level and the rate of inflation can be controlled in a cashless society. In order to analyse this question we study the role of the central bank in such a system.

From the outset it should be stressed that in a cashless society the central bank will have lost the largest part of its *seigniorage* revenue because it has lost its monopoly position in the creation of cash, including settlement balances (i.e. bank deposits held with the central bank to be used for the settlement of payments).[10] Boeschoten and Hebbink (1996) have calculated that the effect of this loss on central bank balance sheets is likely to be large, depending upon the capacity of e-money to substitute for traditional money.

According to these authors, seigniorage revenues largely exceed central bank operating expenses, being an important source of profit for central banks (see Table 13.1). Considering only that prepaid cards would eliminate all banknote denominations up to US$25, seigniorage revenue would diminish between 0.05 per cent and 0.16 per cent of GDP, in different G-10 countries. If all segniorage revenue would be abolished, as we are assuming here, an important source of revenue would disappear and central banks would certainly have to find alternative financing sources.

Table 13.1 Comparison of seigniorage and central bank expenses (1994) (% of GDP)

Country	*Seigniorage**	*Central bank operating expenses*	*Seigniorage reduction (Eliminate all banknote denominations up to US$25†)*
Belgium	0.44	0.17	0.05
Canada	0.31	0.03	0.15
France	0.28	0.13	0.08
Germany	0.52	0.07	0.06
Italy	0.65	0.06	0.05
Japan	0.42	0.06	0.01
Netherlands	0.46	0.06	0.06
Sweden	0.48	0.04	0.10
Switzerland	0.45	0.05	0.05
UK	0.28	0.03	0.14
United States	0.43	0.03	0.14

Source: Boeschoten, W.C. and Hebbink, G.E. (May 1996).

Notes:

* Seigniorage is roughly estimated by multiplying notes and coin outstanding by the long-term rate of interest on government securities.

† Or the equivalent amount in domestic currency.

This immediately raises the issue what the relation of the central bank will be with the government. In a first step we will assume that the central bank is completely independent from the government and will not be subsidized by that government. This also implies that the central bank, like any other bank, cannot make losses in a sustained way. In this context we analyse whether the central bank can use its traditional instruments of monetary policy in a cashless society. We will first study the question of whether the central bank can use open market policies (Section 3.1). In the following section we analyse the issue of whether the central bank can control the short-term interest rate in a cashless society (Section 3.2).

In a second step we will analyse the question of whether the central bank can maintain its status of political independence once it has lost its seigniorage. We will study the new relations between the central bank and the Treasury in a cashless society that arises from this loss of seigniorage. As will be seen, the need to redefine the relation between the central bank and the government in a cashless society follows not only because of the loss of seigniorage but also because the central bank cannot use its traditional instruments of monetary policy without some government support. We will analyse the nature of this support.

3.1 Can the central bank use traditional monetary policy instruments in a cashless society without government support?

The control over the short-term interest rate is crucial in the monetary policies of modern central banks. Even if the latter aims at targeting the money stock, the operational way to achieve the target consists in changing the interest rate that is applied in open market operations and in bank advances. The central bank can also be guided by the new approach to monetary policy, which stresses that the central bank should aim at the inflation rate directly (e.g. in the context of a Taylor-type rule). In this approach the central bank uses the short-term interest rate in response to increases (decreases) in expected inflation rates. Thus, whatever the approach taken by the central bank, the capacity to control the short-term interest rate is of overriding importance. The question then is whether the central bank will be able to control the nominal interest rate in a cashless society. In this section we analyse how the central bank can influence the short-term interest rate using the system of standing facilities and open market operations.

The system of standing facilities is used by many central banks, including the Eurosystem. We first analyse its working in the present cash system, and then go into the question of whether this system can continue to function in a cashless society.

The basic functioning of a system of standing facilities is represented in Figure 13.1. This shows the demand (D) and supply (S) of liquid funds (overnight funds) by commercial banks. The central bank sets its deposit rate at DR and its lending rate at LR. These official rates work as floors and ceilings within which the overnight rate (OR) will be located. A shift in the demand curve, say to D', forces the central bank to intervene and to supply liquidity. The supply curve becomes perfectly elastic when the overnight rate reaches LR. At that point the central bank is willing to supply unlimited amounts of overnight funds.

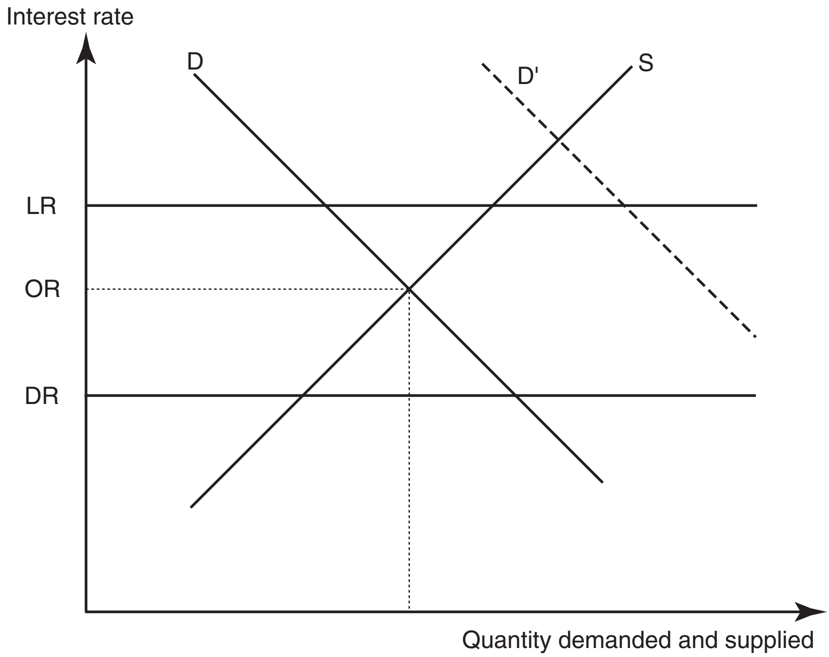


Figure 13.1 The market for overnight balances.

To make this point clearer, consider the mechanics of borrowing at the central bank. When today a bank borrows at the central bank at the announced rate LR, the central bank has no limit in the amount it can lend, for the simple reason that it creates bank reserves 'by a stroke of the pen'. That is, the ultimate means of payments are the liabilities of the central bank. As a result, there is in principle no constraint on the amount the central bank can provide in a lending operation.[11] Surely, when the central bank lends to commercial banks, it will demand collateral, but all this means is that the constraint is on the borrowing institution. The central banker, however, can face any demand of liquidity originating from the banking sector.

The central bank can raise (or lower) LR and DR. This shifts the band upwards (or downwards) forcing the overnight rate to increase. In principle, very little actual intervention in the market will be necessary, because market participants know the commitment and the unlimited capacity of the central bank to intervene at the margin. As a result, when the central bank shifts the band the overnight rate will typically settle within the new band, without the need to intervene in the market.[12] Can this system continue to function in a cashless society? In order to answer this question we continue to assume that the central bank in the future cashless society is independent from the government and that it cannot draw on the state's resources.

Two things that are important for the operation of the standing facility will have changed in a cashless society. First, as stressed earlier, the central bank's liabilities will not be used as cash and as settlement balances any more. Second, the relative size of the central bank balance sheet in a cashless society will have shrunk considerably. Let us develop this last point further. When we will have moved to the point where currency and bank reserves have disappeared, the size of the central bank's balance sheet in comparison with the size of the balance sheets of commercial banks will have declined sharply. To give some perspective, Table 13.2 presents the size of the balance sheet of the Eurosystem (ECB + national central banks in Euroland) and compares this with the size of the balance sheets of the financial institutions (MFIs). It can be seen that today (May 2000) the balance sheet of the Eurosystem represents 8.9 per cent of the consolidated balance sheets of the banking sector in Euroland. This is still a respectable size. When compared to the balance sheet of individual MFIs, the central bank is in most countries still the largest player in the market.

However, it is to be expected that the relative size of the Eurosystem will continue to decline as the use of currency and bank reserves declines. Today these two items represent more than 70 per cent of the total balance sheet of the Eurosystem. There is thus a huge potential for further shrinking of the Eurosystem as a result of the gradual decline in the use of currency and bank reserves. At the same time the financial sector is likely to continue its strong expansion. The last few years it has expanded at rates close to 10 per cent a year. Thus, when we reach the cashless society, the central bank is likely to be much smaller than today. It could become significantly smaller than the big players in the financial markets.

The small size of the central bank in a future cashless society together with the loss of its monopoly position in the supply of cash and of settlement balances has important implications for the operation of standing facilities in a cashless society. Consider again a borrowing operation by a commercial bank. The central bank will not be able then to provide the demanded liquidity by 'a stroke of the pen', because its liabilities will not be the ultimate means of payments anymore (cash and settlement balances). Instead, the central bank will have to go to the market and borrow liquid funds itself, which it will then lend out to the commercial bank. This also implies that the central bank will have to provide collateral. If the borrowing banks are large relative to the central bank, the latter will not have sufficient collateral to make the operation possible. Instead of being an institution that does not face a constraint on its capacity to create liquidity, the central bank

Table 13.2 Consolidated balance sheets of the Eurosystem and of the MFIs in Euroland (May 2000) in billion euros

Eurosystem	*MFIs*	*Eurosystem/MFIs*
1,082.2	12,135.1	8.9%

Source: ECB, Monthly Bulletin, July 2000.

in a cashless society will face the rigours of the market. As a small player in that market it will have great difficulties lending the amount of liquidity that big players will occasionally demand. Thus, the loss of monopoly power in the provision of liquidity combined with the small size will make it hard for the central bank to control the short-term interest rates.

The previous analysis can be made more precise by referring to Figure 13.1. The horizontal lines (floors and ceilings) can be interpreted as infinitely elastic supplies of deposits and loans by the central bank. Take the perfectly elastic supply of loans (the ceiling). This perfect elasticity arises from the fact that the central bank today can supply the loans at the given loan rate LR in unlimited amounts because its liabilities are the ultimate means of payments. In a cashless society where the central bank will have 'to go to the market' in order to raise the liquidity it wants to lend, the supply of loans curve will not be horizontal anymore. We show this in Figure 13.2.

Suppose as before that there is a shift in the demand for liquid funds. As a result at the rate LR there is excess demand equal to AB. The central bank wishing to prevent the short-term interest rate from exceeding LR would have to borrow at market conditions (given by LE, the equilibrium rate, corresponding to the intersection of the D'- and S-lines) and lend the funds to the commercial banks at the rate LR. This would lead to losses for the central bank and profits for the

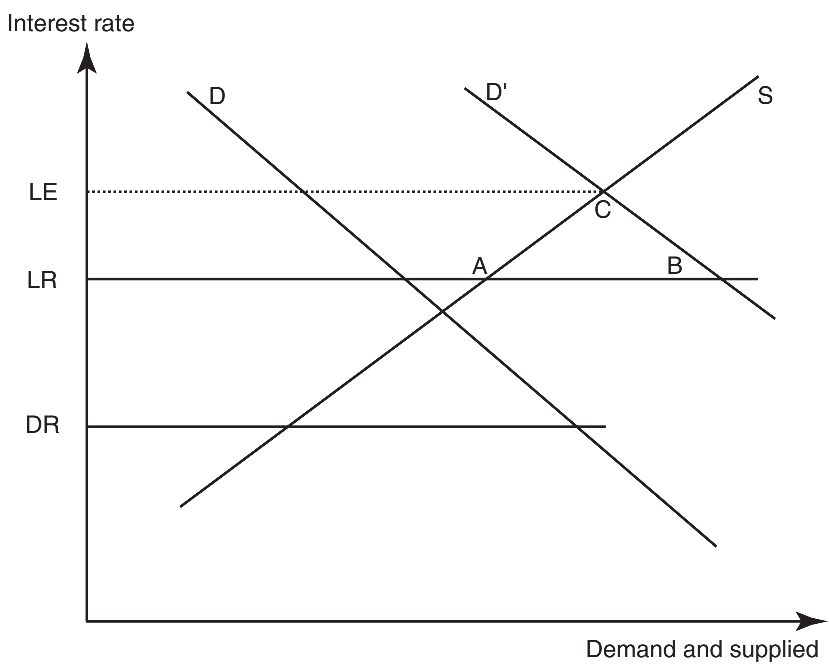


Figure 13.2 The market for overnight balances in a cashless society.

commercial banks. In addition, and more importantly, it would create arbitrage opportunities for the commercial banks that would have a strong incentive to borrow cheaply from the central bank and to lend the funds back to the same central bank. This would lead to unsustainable large losses for the central bank. Without an external support the central bank cannot credibly set up a standing facility in a cashless society. This conclusion is reinforced by the fact that the central bank will have relatively small size. It will, therefore, be impossible for the central bank to keep the short-term interest rate between a horizontal band as in Figure 13.1. All this leads to the conclusion that in a cashless society where the central bank's liabilities are not the ultimate means of payment any more, the central bank will have lost its capacity to regulate the short-term interest rate. Since agents know this the commitment of the central bank to intervene in the overnight market and to be the residual supplier and demander at fixed rates will have no credibility.[13]

The previous analysis about standing facilities can easily be extended to open market operations. The latter will be ineffective in controlling the interest rate for the same reason as in the case of standing facilities. Consider an open market purchase by the central bank. This means that the latter offers to buy Treasury securities held by commercial banks at a price above the market price, so that these are willing to sell the securities to the central bank. As argued earlier, in a cashless society the central bank deposits acquired by the commercial banks are just regular bank deposits. The commercial banks are unlikely to be willing to hold the full amount of these deposits in their asset portfolio. The largest part will be presented for conversion at the central bank, through the private payments system. The central bank will have to liquidate an equivalent amount of assets (Treasury securities). As a result, the initial increase in the assets and liabilities of the central bank is eliminated. The central bank cannot expand the amount of liquidity in the system, except to the extent that other banks are willing to hold the liabilities of the central bank.

It is worth noting that the situation of the central bank in a cashless society is very much comparable to the situation of a central bank of a small country on a fixed exchange rate and perfect capital mobility. Any attempt by this central bank to independently raise the domestic money stock by open market operations would lead to an outflow of reserves, which forces the central bank to bring the money base back to its initial level.

3.2 The central bank and the Treasury in a cashless society

In the previous sections we argued that the central bank in a cashless society will have few instruments available to fulfil its mandate of maintaining price stability. In deriving this conclusion we assumed that the central bank stands alone and cannot draw on the resources of the government. In this section we relax this assumption. We ask the question of how the relationship between the central bank and the Treasury will have to be redefined to allow the central bank to fulfil its mandate for price stability.

Let us consider the case in which the central bank can draw on the resources of the state in a cashless society, i.e. the losses that the central bank incurs are borne by the Treasury. Will this be sufficient to allow the central bank to regain its control over the short-term interest rate? In recent contributions, Goodhart (2000) and Woodford (2000) have claimed that this would be the case.

Goodhart (2000) has made the point that the central bank will maintain its capacity to control the short-term interest rate because it is not a profit maximizing institution, and it is capable of drawing on the resources of the state. As a result it can set the interest rate at a different level from the market rate, knowing that the ensuing losses it incurs will be covered by the Treasury. This gives the central bank the power to move the interest rate in the market.

This reasoning can be criticized on the following grounds. In a society where central banks do not impose minimum reserve requirements, as we are assuming here,[14] deposits held at the central bank for operational reasons tend to be residual.[15] In a cashless society where central banks do not have any specific role in the monetary process any more, there is no apparent reason for banks to hold reserves at the central bank.

Thus, when the central bank sets the loan rate LR (as in Figure 13.2), this means that it will have to borrow the funds itself in the market at the rate LE in order to lend them back at the lower ceiling rate LR. As argued earlier, this will set in motion profitable arbitrage by banks whereby the latter borrow cheaply from the central bank and lend the same funds dearly in the market. The counterpart of this arbitrage are the losses of the central bank. The fact that the central bank now has at its disposal the resources of the state will only make the size of these arbitrage activities larger, without creating added liquidity. As a result, the equilibrium rate remains in LE.[16] If the willingness of the Treasury to bear the losses of the central bank is unlimited, so will the profits of the banks be unlimited. This is clearly unsustainable. We conclude that the willingness of the Treasury to bear the losses of the central bank does not increase infinitely the power of the central bank to set the interest rate. The crucial reason is that the central bank has lost its seigniorage and the power to create the ultimate means of payments (cash and settlement balances). As a result, it cannot affect the amount of liquidity in the system.

A similar conclusion holds as far as the use of open market operations is concerned. Consider again an open market purchase by the central bank in a cashless society. In order to convince commercial banks to sell their Treasury bills the central bank must offer favourable terms, i.e. a premium above the market price. This will be a source of losses for the central bank.[17] The government, however, will foot the bill, but this will not help the central bank to control the stock of liquidity in the system. As argued earlier, the commercial banks that, after the open market operation, have acquired a deposit issued by the central bank will mostly not want to keep these in their books. As a result, these deposits will be presented to the central bank for conversion forcing the latter to liquidate its stock of earning assets, i.e. the Treasury securities it has bought in the first place. The fact that the central bank does not face a profit constraint is irrelevant. What is relevant here is that in a cashless society, the central bank faces the same constraints as the other banks, i.e. it cannot overextend its loans. If it does, it faces deposit withdrawals forcing it to liquidate assets.

In order for the central bank in a cashless society to be able to control the interest rate and the amount of liquidity, one needs more than a commitment by the Treasury to cover the losses of the central bank. In addition, the Treasury must be willing to supply potentially large amounts of risk-free securities (bonds) to the central bank. There are two reasons why this is necessary. First, large holdings of Treasury securities by the central bank would make the central bank a big player in the financial markets. Put differently, the shrinking balance sheet of the central bank on the road to a cashless society would be 'blown up' by Treasury securities.

Second, and most importantly, the capacity of the central bank to obtain Treasury securities is essential to allow it to use its traditional instruments of monetary control in a cashless society. Take the standing facilities. We have seen that in a cashless society the central bank must borrow funds in the market when it faces a demand for loans at the loan rate LR. In order to do so, it has to provide collateral. This collateral can only come from Treasury securities it is holding, or that it can obtain from the Treasury. In addition, by supplying new Treasury securities in the market, the central bank is increasing the amount of liquidity in the system, thereby affecting the interest rate.

In a similar way, the central banks will have to be able to draw on a pool of Treasury securities to sustain an open market purchase. This allows the central bank to confront withdrawals after an open market purchase. Thus, in a sense, in a cashless society, Treasury securities become the ultimate means of payments. By being able to draw on an automatic credit line with the Treasury, the central bank can effectively control the total amount of liquidity.

There is a danger associated with this required transformation of the role of the central bank in a cashless society. The danger is that the central bank becomes an agent of the Treasury (i.e. the government), so that it loses its independence. We have learned from the monetary history of the last centuries that central banks that are dependent on political institutions often fail in stabilizing the price level. As a result, many countries in the world have transformed their central banks into institutions that are independent from the politicians. There is a large consensus today that this transformation is beneficial and that it is necessary to maintain price stability.

The paradox therefore is that the movement towards a cashless society may in fact increase the power of the state on monetary affairs again. Thus, the challenge becomes how to devise institutions that will give the means to the central bank to play its role in stabilizing the price level, without increasing the intrusion of politicians.[18] Certainly such institutional changes are possible. They will require some creative thinking, however.

4 A new role for the central bank in a cashless society

In the previous section we argued that the central bank will lose both its traditional instruments of monetary policy and the largest part of its revenues in a cashless society. It therefore risks becoming very much dependent on the

Treasury, both as a means to strengthen the use of its traditional instruments of policy and as a means of obtaining revenue.

There is, however, another avenue for the central bank (or more broadly the monetary authorities). This consists in strengthening the supervisory role of the monetary authorities. (Note that we abstract here from the fact that in some countries supervision is done by an institution other than the central bank. In the present context we will assume that the central bank is also the supervisor.)

At the core of the central banks' responsibility in a cashless society will be supervision of the money-issuing institutions. By supervision we mean the control of any activity which might interfere with monetary variables or with the sound functioning of the monetary system. In a cashless society this supervision should be strengthened in two directions.

First, the central bank will have to expand its supervisory control of any money issuance, independently of its type or origin. It will then be possible to implement a system where the central bank certifies the quality of the issue of private money. This will apply to both traditional private money (such as bank deposits) to new forms of money such as e-money. In such a system, if a firm has enough assets, which allow for money issuing, the central bank will 'print' its 'logo' on this money, assuring the underlying quality of the asset. The central bank will therefore perform a rating activity, which will give legal tender characteristics to private money, including e-money.

Second, the central bank as a supervisor should not only use microeconomic criteria to certify the quality of the money-issuing institution. In a cashless society it should also use macroeconomic criteria to guarantee the quality of the money issue, including e-money. The use of macroeconomic criteria will be crucial precisely because the central bank will have lost most of its traditional instruments to influence macroeconomic variables such as the money stock and the interest rate. Using macroeconomic criteria in the supervision will then substitute for the lack of direct control over macroeconomic conditions in a cashless society.

Let us give an example to make this point clear. Consider the quality of mortgage loans. In a traditional microeconomic approach to supervision, the central bank (or the institution responsible for supervision) evaluates the quality of the banks' loan portfolio by analysing the structure of the bank's balance sheet, the amount of equity issued, the value of the collateral, etc. The macroeconomic environment in which these loans are issued usually does not play an important role.[19] The use of macroeconomic indicators to evaluate the quality of the loan portfolio would include an analysis of the degree of asset inflation in the housing market, the risk associated with interest rate movements, the state of the business cycle. When these macroeconomic indicators point towards overheating, and thus to a risk of asset inflation, the supervisor could strengthen the criteria for providing its positive rating. For example, during a boom in economic activity, the supervisor could raise the capital adequacy ratio, or alternatively it could increase the collateral banks are required to use in extending loans. During a recession the supervisor could then lower the capital adequacy ratio and/or lower the required collateral for extending bank loans. Used in such a way, supervision could in fact become an instrument to affect macroeconomic conditions in a counter-cyclical way. Such a counter-cyclical supervision would give the monetary authority an instrument to stabilize the price level in an environment where it has lost many of its traditional monetary policy instruments.

In this expanded role for supervision the use of legal reserve requirements will be of great importance. It will have to be applied to non-traditional types of money, such as e-money. In practice this means that a certain percentage of 'high quality' e-money should be deposited at the central bank. This would not necessarily make supervised firms less competitive. It would depend upon the remuneration rate attached to these deposits. By remunerating deposits with an interest rate below but close to the market rate, the central banks can make sure that the losses of competitiveness of the money-issuing institutions are limited. The price these institutions would have to pay in the form of a somewhat lower return on short-term assets than the market return could be considered to be the price to pay for a permanent and updated rating service.

Legal reserve requirements can also be used as an instrument of control of the supply of money, including e-money. One should be aware, however, that as an instrument of monetary control, legal reserve requirements have such a large impact, that they have to be used very carefully. Consequently, their flexibility is quite limited, especially in the absence of the other traditional instruments of monetary control.

It is useful to look at the present-day use of legal reserve requirements. Table 13.3 shows the recent use of reserve requirements in a number of industrial countries and in Euroland. It can be seen that reserves are very small.

In Germany, it used to be more important, but since the start of EMU, the importance of reserve requirements in that country has declined. In 1997, the Dutch Central Bank used reserve requirements primarily for liquidity management functions. In order to maximize the flexibility of the requirements, authorities flexibly adjusted the level of the ratio and the length of the maintenance period in order to meet the changing profile of the autonomous creation of liquidity. However, subsequently, this policy was slowly substituted by more market-oriented instruments.[20]

In the Eurosystem, the reserve base includes 'deposits', 'debt securities issued'

Table 13.3 Reserve deposits held at the central banks as a share of total bank liabilities (year average of end-month observations, in percentages)

	1980	*1985*	*1988*	*1991*
United States	1.6	0.8	1.0	0.6
Japan	1.6	1.1	1.0	1.0
Switzerland	4.0	3.1	1.7	0.7
UK	0.3	0.1	0.1	0.1
Germany	7.2	5.5	5.5	0.1

Source: Kasman, B. (1992) 'A comparison of monetary policy operating procedures in six industrial countries', *Federal Reserve of New York*, vol. 17, no. 2, Summer and 'money market paper with maturities inferior to two years'. The minimum reserve requirement ratio is only 2 per cent of reserve base. Thus, the use of legal reserve requirement today is very limited. In order for reserve requirements to become an effective instrument of monetary control, they would have to be expanded significantly, so that they could be used both as an instrument of monetary contraction and of monetary expansion. With reserve requirements of 1 to 2 per cent of the reserve base, significant relaxation of monetary policies are impossible.

It is quite unlikely that in a cashless society, which will also be characterized by fierce competition and internationalization of the financial markets, the monetary authorities of individual countries can significantly increases legal reserve requirements. As a result, their use as an instrument of monetary control will remain limited.

5 Conclusion

In this paper we have analysed how monetary policies will be affected in a cashless society. Our main conclusions are that the central bank will lose its traditional instruments of monetary policy. Standing facilities and open market operations will become ineffective as instruments to control the interest rate and the money stock. This is problematic because in a cashless society where all the money is privately supplied, there is no clear and reliable mechanism that ties down the price level.

We also argued that this leads to two possible avenues for the future role of the central bank. In the first one the central bank becomes very dependent on the Treasury, both as a means of obtaining revenues and as a way of maintaining some effectiveness for its traditional instruments of monetary policies. This road is not without dangers because it would imply a return to a system of political dependence of the central banks.

Another avenue consists in redefining the role of the central bank. This consists in strengthening the supervisory role of the monetary authority. This strengthening would include the use of macroeconomic indicators in the control of the quality of the loan portfolios of the money-issuing institutions. It would lead the central bank (supervisor) to design its supervision in a counter-cyclical way. It also implies that the supervision should be expanded to the issuers of e-money.

Notes

1 Recently, strong doubts about the speedy disappearance of cash have been formulated by Goodhart (2000) and Woodford (2000). In this paper we do not go into this problem. We take the view that since the advent of a cashless society in the foreseeable future cannot be excluded it is interesting to analyse its implications for monetary policies.

2 This definition makes our cashless society very different from the free banking system that has been analysed by, among others, Hayek (1978) and Klein (1974). In the free banking system each bank issues its own currency and/or deposits with different units of account. In such a system there is a free floating exchange rate between the different

moneys even within the same country. It is conceivable that e-money would be of the same kind, i.e. each issuer uses a different unit of account. We do not go into this analysis here.

3 For a more rigorous treatment see Patinkin, (1965).

4 See Patinkin (1965) for a generalization. Patinkin's generalization is based on dichotomizing the economy. In equilibrium, demand and supply in the real sector of the economy depend on relative prices only. This equilibrium is unaffected by the aggregate price level and the money stock. The latter only matter for the equilibrium in the money market.

5 The fact that they have the capacity to do so does not mean that they always do. The twentieth century is replete with histories of spectacular failures that have led to hyperinflation.

6 For an analysis see von Hagen (1999) and Neumann (1999).

7 The latter authors find that this is what the Bundesbank was doing *de facto*. See also Bernanke and Mihov (1997) and Clarida and Gertler (1996).

8 There are dissenting views, however. The most notorious one was formulated by Hayek (1978). See also Klein (1974). These authors analysed the problem in the context of free banking, i.e. a system in which each deposit-issuing institution creates its own unit of account. Here we consider a system in which the unit of account is a collective good supplied by the state.

9 The ECB's minimum reserve system has as two main functions: '] … [stabilization of money market interest rates] … [' and '] … [creation or enlargement of a structural liquidity shortage] … [', according to 'The Single Monetary Policy in Stage Three – General Documentation on ESCB Monetary Policy Instruments and Procedures', ECB, September 1998.

10 This does not preclude that the central bank issues its own deposits that are used as money. The central bank could then make money like any other private bank.

11 This is also the basis of the capacity of the central bank to be a lender of last resort.

12 See Guthrie and Wright (2000), who analyse a model in which the announcement of a particular target rate is sufficient to stabilise the interest rate without any need for intervention. The key for this result is the knowledge that the central bank has the capacity for unlimited intervention.

13 The reader will recognize that the problem of the central bank in a cashless society is comparable with the problem of the central bank pegging a fixed exchange rate. In the absence of sufficient reserves this commitment has low credibility and will be fragile. It will often lead to speculative crises. We came to the same conclusion when we discussed the use of open market operations in a cashless society.

14 We drop this assumption in Section 4.

15 According to Borio (1997), in countries such as the UK, where the reserve ratio was very small (below 0.5 per cent), deposits tend to be very small (below 0.5 per cent of the GDP), showing that even in a traditional money supply environment banks do not hold deposits with operational goals, namely if they are not remunerated.

16 Note that this is where the cashless society differs from the present monetary regime. In the latter the lending by the central bank adds liquidity, and thereby affects the equilibrium in the market.

17 Note that this is also the case today in a cash society. The central bank must also offer favourable terms. The losses it makes on open market operations are more than offset by the large profits resulting from the fact that most of the liabilities of the central banks do not pay interest (the seigniorage gain). The latter is absent in a cashless society.

18 The prediction made by Goodhart (2000) that the central bank would be the same institution as today, but instead of making monopoly profits, would make (potentially) large losses that would be covered by the state, is not very comforting. It is difficult to see how an institution that year in year out must be subsidized by the government would remain independent from the latter.

19 We do not want to imply that present-day supervisors do not use macroeconomic analysis. Some undoubtedly do. We argue that in a cashless society this will have to be done more systematically.

20 See Borio (1997).

References

Bernanke, B. and Mihov, I. (1997) 'What does the Bundesbank target?', *European Economic Review* 41: 1025–52.

BIS (1996) *Implications for Central Banks of the Development of Electronic Money*, October 1996.

Boeschoten, W.C. and Hebbink, G.E. (1996) 'Electronic money, currency demand and seigniorage loss in G-10 countries', *De Netherlandsche Bank Staff Report*, May.

Borio, C.E.V. (1997) 'The implementation of monetary policy in industrial countries: a survey', BIS Economic Papers no. 47, July.

Clarida, R. and Gertler, M. (1996) 'How does the Bundesbank conduct monetary policy?', NBER Working Paper no. 5581.

Clarida, R., Gali, J. and Gertler, M. (1997) 'Monetary policy rules in practice: some international evidence, CEPR Discussion Paper no. 1750.

Clarida, R., Gali, J. and Gertler, M. (1999) 'The science of monetary policy: a new Keynesian perspective', CEPR Discussion Paper no. 2139.

Fama, E. (1980) 'Banking in the theory of finance', *Journal of Monetary Economics* 6: 39–57.

Fisher, I. (1896) 'Appreciation and interest', *AEA Publications* 3(11), August: 331–442.

Friedman, B. (1999) 'The future of monetary policy: the central bank as an army with only a signal corps', *International Finance* 2(3), November: 321–38.

Goodhart, C. (2000) 'Can central banking survive the IT revolution?', paper prepared for the conference *The Future of Monetary Policy*, held at the World Bank, 11 July 2000.

Guthrie, G. and Wright, J. (2000) 'Open mouth operations', *Journal of Monetary Economics* 46: 489–516.

Hayek, F. (1978) *Denationalising Money: the Argument Refined.* London: Institute of Economic Affairs.

Klein, B. (1974) 'The competitive supply of money', *Journal of Money Credit and Banking* 6, pp. 423–53, November.

Neumann, M. (1999) 'Monetary stability: threat and proven response', in Deutsche Bundesbank, *Fifty Years of the Deutsche Mark*. Oxford: Oxford University Press.

Patinkin, D. (1965) *Money, Interest and Prices* (2nd edition). New York: Harper and Row.

Taylor, J. (1993) 'Discretion versus policy rules in practice', *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.

von Hagen, J. (1999) 'A new approach to monetary policy (1972–8)', in Deutsche Bundesbank, *Fifty Years of the Deutsche Mark*. Oxford: Oxford University Press.

Woodford, M. (2000) 'Monetary policy in a world without money', paper prepared for the conference *The Future of Monetary Policy*, held at the World Bank, 11 July 2000.

Acknowledgements

The views expressed in this paper are those of the authors. They should not be interpreted as reflecting the views of the Banco de Portugal. The financial support of the Flemish Science Foundation (FWO: no. G.0331.01) is gratefully acknowledged. We are grateful for the comments of Stefan Schmitz and the participants of the SUERF meeting.

14 The dynamics of payment networks*

Gottfried Leibbrandt

All the major developed countries have created very different retail payment systems. These differences continue to exist, even though all countries have had access to the same payment technologies during the past thirty years.

This paper argues that this is due to the presence of network externalities. The resulting excess inertia may or may not be overcome, depending on the market structure. A model is introduced to analyse the role of expected share of payment technologies and of international or inter-regional linkages.

Application of the model suggests that the differences in payment systems will persist. Given the current market structure, even new technologies, like mobile payments, are likely to be implemented in different and incompatible ways by different countries.

1 Significant and persisting differences between countries

The days when cash was the dominant way to settle transactions are long gone. Non-cash instruments such as cheques, giro transfers and cards are widely used alternatives. However, there are *significant differences in the instrument mix across countries*. Figure 14.1 gives an overview of the mix of payment instruments for eight major countries.

Three major differences stand out. First, the number of non-cash transactions per person per year is very high in the USA (360) and very low in Italy (40), with the other six countries at an intermediate level of around 150 to 175. Second, the 'Anglo-Saxon countries' and France rely on cheques, where the Benelux and Germany use ACH/giro transfers; here Automated Clearing House (ACH) and Giro represent technologies where a customer instructs his own bank to transfer money, instead of giving a cheque to another person. Third, the Anglo-Saxon countries use credit cards extensively, while France and the Benelux rely almost exclusively on debit cards; Canada uses both, and Italy and Germany use neither.

In addition, there seems to be little convergence between countries. In the USA rumours of the death of cheques are greatly exaggerated: the use of

* Paper presented at 23rd SUERF Colloquium, Brussel, 27 October 2001

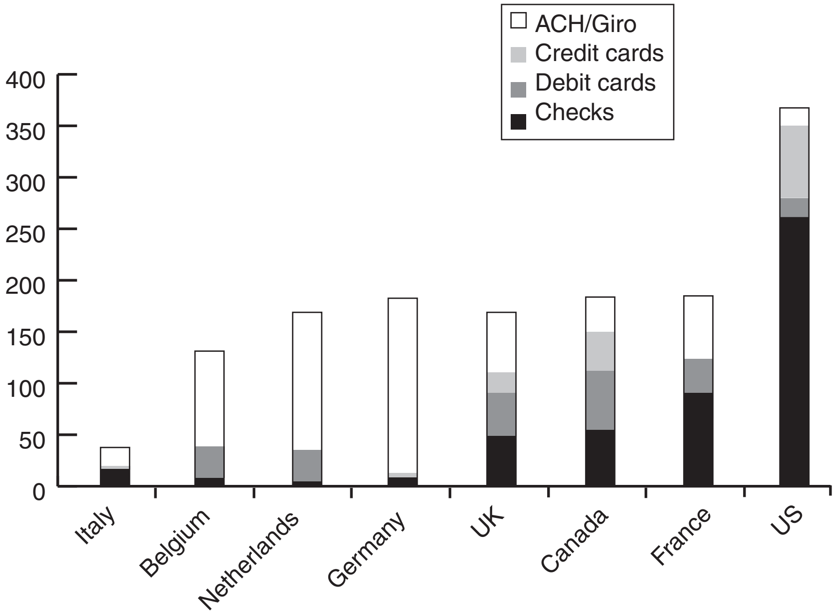


Figure 14.1 Non-cash payment instruments across countries (1999).
Source: BIS (2001).[1]

cheques keeps rising, and new electronic payment instruments may even enhance their use – 'more cheques are written to pay credit card bills, than are replaced at the point of sale' according to one estimate.[2] Similarly, in continental Europe credit card usage shows no intention of growing to Anglo-Saxon levels.

In summary, there are large persisting differences in the usage of various payment instruments across countries. This paper aims to address the following two questions:

1 What causes these differences?
2 Are these differences likely to persist or disappear in the near to medium future?

The answers to these questions are relevant for several reasons, which are listed below.

1 *Significant cost differences:* In general, electronic instruments are estimated to be 55 to 70 per cent cheaper than paper-based instruments. For example, the social cost of a US cheque is estimated at \$2.90 compared to \$1.30 for an ACH/Giro payment.[3] According to one estimate, the USA would save \$91 billion per year, or 1.25 per cent of GDP, if it would switch from cheques to ACH/giro transfers.[4] Even for countries that do not rely on cheques, the savings could be substantial: Norway could save 0.6 per cent of GDP if it were to go from paper to electronic payments.[5] Similarly, there are indications that the

social costs of a credit card transaction are almost twice those of a debit card payment: \$0.88 and \$0.47 respectively.[6]

2 *Implications for competitive environment:* While this paper examines the potential impact of industry structure (essentially concentration) on the choice of payment infrastructure, the reverse relationship is also relevant. For example, credit cards can be (and are) easily issued by specialist players (MBNA, Citi); debit cards, by contrast, have to be issued by the institution that holds the basic salary account.[7] Also, it has been argued that the giro system, with its standing orders and direct debits, increases consumer switching costs.[8] Thus, taken together, the reliance on debit cards and giro systems (as observed in many European countries) may well increase the grip of traditional banks on their customers.

3 *Likelihood of a single euro payment infrastructure:* In just a few months, much of Europe will switch to a single physical currency. This is not the same as one payment infrastructure. Indeed, if the current differences persist, cross-border payments will probably remain much more expensive than local payments.

4 *Insight into standardization process:* Finally, understanding the mechanism of standardization in a multicountry context may lead to insights about the likely adoption of future payment technologies.

The remainder of this paper is structured as follows: Section 2 gives an overview of the existing literature on usage of payment instruments. Section 3 introduces the basic frameworks from the 'network literature'. Section 4 describes a model to apply these frameworks, while Sections 5 and 6 apply this model to resolve each of the two question raised above. Finally Section 7 summarizes the findings and implications.

2 Traditional payment literature offers only partial explanation

There is of course a rich literature on the trade-off between holding cash and non-cash money, starting with Baumol's inventory model, where a consumer makes a trade-off between the cost of a trip to the bank to get cash and foregone interest revenue on cash.[9]

However, there are relatively few theoretical models that explain the choice of various non-cash instruments. Given the many variables involved in these choices they tend to focus on one aspect, such as cheque float, or the difference in processing cost between credit cards and stored-value (smart) cards.[10] However, these models cannot be readily used to analyse or explain cross-country differences, because they relate the choice of payment instrument to factors like transaction size, inherent characteristics of the instrument, i.e. factors that are not obviously different across countries. This is an important gap: any model that wants to explain differences across countries will have to look at country-specific factors, like regulation, industry structure or just seemingly random events in a country's development. This paper aims to fill that gap. As such it tries to build on the work of the few authors that *have* sought to trace payment differences to country-specific factors. The most relevant findings of these authors are discussed below.

The comprehensive empirical survey of cross-country differences by Humphrey, Pulley and Vesala (1996) concludes that country idiosyncrasies seem to play a bigger role than factors like *per capita* income and crime rates. The study analyses data on five instruments (cheques, credit cards, debit cards, paper giro and electronic giro) over seven years and across ten countries.[11]

In their 2000 article, the same authors analyse 'America's love affair with cheques'. In the best romantic tradition, this love affair appears to be both irrational (i.e. socially costly) and persistent. Several explanation are offered:

1 *Perverse pricing:* The writer of a cheque gets the benefits (primarily float), while the recipient bears the cost. This argument is elaborated in Humphrey and Berger (1991) and disputed by Wells (1996), who argues that the average float has declined significantly, from \$1.04 per cheque in 1987 to \$0.09 in 1993, and is anyway insignificant for the average consumer cheque.[12] A second pricing perversity is the fact that few banks charge consumers per cheque. However, research finds the impact of prices on consumer cheque use to be limited.[13] Finally Shy (2001) blames the subsidy of cheque clearing by the Fed. Since the Fed only performs a very small part of cheque clearing, representing in the order of \$0.05 per cheque, this argument too remains unconvincing.[14]
2 *Sunk investments in processing cheques:* Two findings suggest that this effect may be smaller than some authors assume. First, scale economies in cheque processing are exhausted at fairly low volumes. Second, over the last decade, studies have not found evidence of significant technological change.[15] Both findings suggest there has been little economic need to invest in large-scale infrastructure for processing cheques.
3 *Fragmentation and strong anti-trust regulation:* Any individual bank has little incentive to move to electronic transfers, while anti-trust regulation complicates coordination among banks. This is a very likely explanation (it is essentially the 'network effect' examined in the next section).
4 *Convenience:* As Alan Greenspan recently put it: 'currency and to a large degree cheques are currently perceived to offer significant advantages in privacy over electronic payment systems, (…). Perhaps an even more important dimension influencing our behavior regarding money and payments is convenience'.[16] Several studies have confirmed that non-cash instruments are more convenient for consumers for larger expenditures, while cash is more convenient for small purchases.[17] Again, these are very plausible and powerful explanations, but by themselves they do not explain the differences. Why do Americans find it more convenient to use cheques compared to consumers in other countries?

To summarize, in spite of several interesting approaches, the significant differences can only be partially explained. It is quite plausible that regulation and industry structure play an important role, but the specific relationship remains to be explored.

3 Network economics are very relevant in payment systems

Many payment systems exhibit what is popularly called 'chicken and egg problems'. For example, consumers will only use a credit card if it is widely accepted, and merchants will only accept the instrument if many consumers carry a card. In economic literature this effect is called a *network externality*.[18]

Over the past fifteen years a wide body of literature has analysed these effects.[19] According to a 1996 survey: 'the economics of compatibility have now evolved into a well-recognized subfield of industrial organization. (…) The mechanisms which can lead to excess or insufficient standardization through product design and/or adoption decisions are well-known'.[20]

In general it is found that network externalities can lead to strong *increasing returns*, popularly known as 'winner takes all': once a standard reaches critical mass, increased usage will increase its value, which further increases it market share at the expense of competing standards. This in turn may cause '*path dependence*' and '*excess inertia*'. Path dependence means that equilibrium outcome can be influenced by minor coincidences in the past instead of being pre-determined by fundamental forces.[21] Excess inertia refers to the fact that society may adopt inferior standards, or networks that are smaller than the social optimum.

Network effects have been empirically analysed in areas like fax machines, DVDs, ATMs and VCRs.[22] And several authors have described examples where inferior solutions did prevail in nuclear technology and typewriter keyboard layout (QWERTY).[23]

Quite a few authors have argued that network effects are relevant in the choice of payment method. Perhaps most fundamentally, it has been argued that fiat money itself is a network good, deriving its value from the fact that a critical mass of participants in an economy accepts it as payment.[24]

In addition, a wide variety of theoretical models have been developed. In a payment system context, two issues in particular have been explored:

1. *Compatibility decisions*: Should a bank or a consortium link its network (ATM, debit card etc.) to that of other players, and should it be allowed to do so from an anti-trust perspective? and
2. *Fixing of interchange fees*: Should banks or consortia be allowed to set fixed interbank prices for the use of these networks?[25]

However, most models assume infinitely small or equal-sized players operating in a single economy. As a result, the impact of player size and the interaction between semi-connected regions or countries have received little attention. There are, however, strong indications that both factors are very relevant. For example, it is known that technological development can be quite local. The few models that explore such spatial differences in technology development indicate that increasing returns may enhance initial differences between countries, leading to technological divergence.[26]

4 A simple model

Assume a number of banks that each face the decision on whether to adopt a payment technology g, which is different from their current technology f. They each have a set of customers that make payment transactions using either f or g.

All banks start from a situation where they are using technology f. Without loss of generality I normalize the cost and benefits of technology f to zero. I also normalize the number of transactions per customer to one.

The use of g will cost a bank a fixed cost per customer, denoted c.[27] Let b be the benefits per transaction for the bank. These could, for example, be lower costs per transaction as a result of using g instead of f.[28] However these benefits are only realized if the banks of *both* customers involved in the transaction support technology g (this could be the same bank). If not, the old technology f has to be used.

Assume that customers initiate their transactions randomly with other customers, i.e. they don't have a preference for transactions with customers of their own bank.[29] If bank i, with market share s_i, is the only one using g, then it can use that technology for a fraction s_i of its transactions. If the share of other banks that have adopted g is s_g, then bank i can use g for a fraction $(s_i + s_g)$ of all its transactions. Assume $b > c$, i.e. the new technology g is profitable if *all* banks would adopt it.

I also assume that: (1) c and b accrue to the bank of payor (the customer that initiates the transaction), (2) the volume of the transactions made by consumer is exogenous.[30] Finally I assume that banks will keep the benefits for themselves and not pass them on to the market through lower prices.[31] This implies that market shares are not affected by the adoption of g. While this seems a strong assumption, it is supported by analysis that indicates that this is indeed generally the best policy for banks; it can be shown that in a 'hotelling' duopoly any action by banks to use the benefits of g to gain market share through lower prices will lower profits for *both* banks.[32]

I will now use the concept of Nash-equilibrium to define possible outcomes of decision making by banks. A Nash-equilibrium is a situation where no player can increase his utility through *unilateral* action. In particular I am interested in situations where multiple Nash-equilibriums may occur, with at least one such equilibrium being suboptimal from a total welfare point of view.[33] I will call such a situation *lock-in*: the joint welfare of all players increases if they move to another equilibrium, but no player has sufficient individual incentive to do so. As the following proposition states, this is more likely to occur in a fragmented market.

> *Proposition 1. 'Fragmentation increases the risk of lock-in': a welfare suboptimal Nash-equilibrium exists if and only if the market share of the largest player is lower than the cost/benefit ratio of the new technology* g.

The formal proof of this and all subsequent propositions is given in Appendix 14A, but the crux of Proposition 1 is that lock-in can occur only if no individual player gains sufficiently from *unilaterally* switching to g. The largest player (whose share is noted as s_1) has most to gain, and he will not switch if $s_1 b - c < 0$, which is the same as $s_1 < c/b$. On the other hand, if the market leader has sufficient share and switches, then all other players will follow suit, because each player can now use the new technology g for a share of its transactions equal to s_1, plus its own share.

Figure 14.2 illustrates where lock-in may occur. The horizontal axis is s_1 the market share of the largest bank, while the vertical axis depicts the cost/benefit ratio of g. Lock-in can occur only in the upper-left triangle. For example, if g has a cost/benefit ratio of 50 per cent, the necessary critical mass is 50 per cent. So even in this case where the potential economics are quite favorable (benefits are double the costs if everybody adopts g), a bank will only unilaterally adopt g if it controls more than half the market. This is an illustration of the excess inertia mentioned earlier.

Now of course this analysis suffers from the general critique of Nash-equilibriums: players can coordinate their actions to overcome lock-in. However, several things can prevent this from happening in practice; for example, different benefits and costs for different players, uncertainty with regard to benefits and costs, 'gaming' where a coalition with critical market share keeps other players out.

4.1 Effect of upgrading an old technology

This excess inertia can be further enhanced by the existence of an upgrade F to the existing technology f. The effect of such upgrading is to *raise* the level of participation that is needed for the adoption of g.

> *Proposition 2. 'Upgrades increase lock-in': The availability of upgrades increases the number of situations where welfare suboptimal equilibriums occur.*

The crux here is that F is compatible with f, so it can be used for *all* transactions, where g can only be used if *both* parties have adopted g. So even small players will adopt F unilaterally. At the same time adoption of F reduces the benefits of g

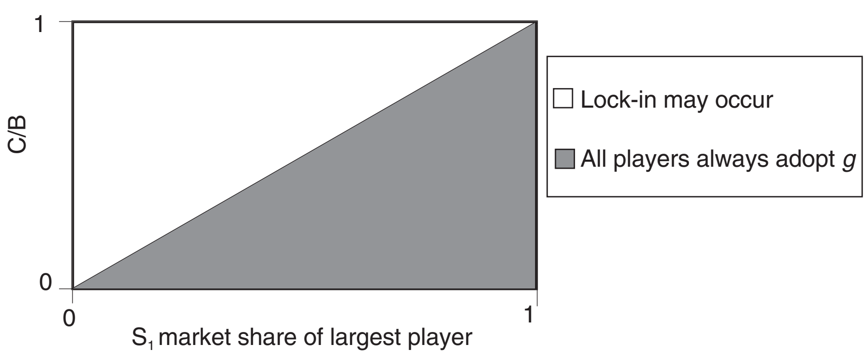


Figure 14.2 Critical mass needed to adopt g.

over f, thereby increasing the market share needed to unilaterally adopt g: as b decreases, c/b goes up, and so does the critical market share needed for a unilateral switch.

Thus a country (or region) can get 'locked in' to an old technology f. Perhaps the most famous example of such 'extended play' of an old technology is the clipper sailing ship of the late nineteenth century, where innovations like steel hulls and extra masts (some of the later clippers had seven masts) extended the life of sailing technology by as much as thirty years.[34]

4.2 Autarkic transaction patterns

So far I have assumed that customers initiate their transactions randomly with other customers, i.e. they don't have a preference for transactions with customers of their own bank. Thus for each bank i, the share of transactions *within* the bank is equal to its market share s_i, and the share of transactions with customers of other banks, or *interbank* transactions, is equal to $1 - s_i$.

I now assume instead that for each bank i, the share of interbank transactions is proportionally lower. Let q_i denote the share of interbank transactions for bank i, then $q_i = \delta(1 - s_i)$. Here $\delta = 0$ corresponds to complete 'autarky' (no interbank traffic) and $\delta = 1$ corresponds to random traffic across banks (the assumption used earlier).

In practice, *within* a country this preference for transactions with customers of the same bank appears to be limited; for example in The Netherlands, δ is close to 1. However, the effect is quite significant *across* countries. Here the basic unit is not banks but countries. In that case s_i is the share of each country in the total, and q_i is the share of international transactions for country i. There is a very strong tendency for people to transact with people in their own country. As will be shown in Section 5, for traffic across (European) countries δ is 0.02. Since for a small country $\delta \approx q_i$, this means that only 2 per cent of all transactions of a small country are cross-border, while for larger players it will be even less.

As the two next propositions assert, semi-autarky is a double-edged sword: it promotes innovation, but also fosters the adoption of different versions of the new technology that may be incompatible with each other.

> *Proposition 3. 'Autarky promotes innovation': A low δ will decrease the occurrence of lock-in into an inferior technology.*

Intuitively, autarky increases the share of 'in-house' transactions, making it more attractive to unilaterally adopt g.

4.3 Incompatible versions of a network technology

With many new technologies, there is no natural or common standard at the outset. Instead, many different, often incompatible, versions are available. Sometimes a common standard emerges (as in VHS *vs*. Betamax and V2000 video formats), but often multiple standards persist (as in metric *vs*. US and UK measurements, or lefthand *vs*. righthand driving).

As the next proposition states, the occurrence of such multiple incompatible standards is related to the degree of autarky.

> *Proposition 4. 'Autarky breeds diversity': (a) Multiple incompatible versions of a network technology* g *can form a Nash-equilibrium if and only if* $\delta < 1$ *and there are positive costs to switching between versions of* g*; (b) the probability of the adoption of incompatible versions increases as* δ *decreases; (c) if autarkic players (*$\delta < 1$*) adopt incompatible versions, and* δ *subsequently rises to 1, the players will always pass through a phase where no individual player will migrate to a common technology, while a switch to the standard of the largest player would enhance overall welfare.*

The intuition here is that autarky reduces the amount of interbank transactions, and thus the benefits of adopting a common standard. If there are positive costs to switching to a common version, then these switching costs may outweigh the extra benefits of being able to use *g* also for interbank transactions.

Propositions 3 and 4 have important implications. In many cases banks (or regions or countries) start out as more or less autarkic. For most of the twentieth century banks were generally local, and so were transaction patterns; national banks and national companies did not emerge until the latter part of the century. As will be shown in Section 5, the internationalization of transaction patters (and banking) has only barely begun. This initial autarky fosters the quick adoption of new network technologies. However, due to this quick adoption, the technologies are often in an early stage, without a clear standard. Because of the autarky the welfare loss of incompatible versions is small, so players may prefer speed over compatibility. But if δ subsequently rises (i.e. the banks or countries become more interlinked) the welfare loss may become significant. Even then, it may not be beneficial to migrate to a common standard from an overall welfare point of view, given the migration costs. As δ continues to rise, a point is reached where lack of coordination may become a problem: no individual bank moves, but overall welfare would increase if everybody adopted a common standard, even after allowing for migration costs.

Propositions 3 and 4 assume semi-autarkic players; these could be either banks or whole countries. Since it is assumed that each player acts as a unit, the model can only be applied to whole countries, if all banks in a country jointly take payment technology decisions. The more complex case of *individual* banks taking separate decisions *within* semi-autarkic countries is treated in Appendix 14B. It broadly leads to the same conclusions as formulated in Propositions 3 and 4, but adds industry concentration within countries as a factor: if such concentration is high, a country will move faster from *f* to *g*, but it is more likely to adopt a version of *g* that is incompatible with other countries.

4.4 Model summary

The above model assumes that the returns to an individual player of adopting a new payment technology depend on the market share of the players that use it.

The model leads to the following implications:

1 Without cooperation, a market consisting of several smaller players may fail to adopt a new payment technology that would be welfare enhancing if adopted by everyone.
2 The availability of upgrades to the old technology may raise the level of cooperation that is required to adopt a new technology, leading to 'lock-in' into a certain technology path.
3 In a context of somewhat autarkic banks (or countries), the level of interaction (or interconnection) is key. A low level of interaction makes it easier for any bank to adopt the new technology on its own. On the other hand, such low interaction may lead to a diversity of technologies being adopted by various banks (or countries).
4 Finally, if each of the countries in turn comprises individual players, the local industry structure becomes relevant: concentration will foster the adoption of *g*, but it also increases the likelihood of incompatible versions.

5 Application to cheques versus ACH/giro

Can this model explain why some countries have adopted giro clearing mechanisms while others continue to rely on cheques?

5.1 How giro was introduced

In most Western economies giro clearing was introduced in some form during the period 1890–1920.[35] Generally this was done through public initiative, either by the state (setting up postal giro systems, using the post offices for access to accounts), or by large cities in the form of municipal giro systems. This system required participants (mostly businesses and public institutions) to maintain giro-accounts. It was only during the expansion of banking into the mass customer segment in the late 1950s and 1960s that banks in most European countries either joined these giro systems or established their own giro-transfer systems to facilitate transfers between bank accounts.

For example in The Netherlands, the Postgiro was established in 1918. This public institution made transfers between customers who had to maintain an account for that purpose. In addition each major Dutch bank had developed its own internal payment transfer system for transfers between customers. However each bank had its own forms and account numbering system, and transfers between customers of different banks were cumbersome.

In 1967 the commercial and agricultural banks decided to establish their own joint giro clearing exchange, in addition to the existing Postgiro.[36] At the time the commercial banks jointly held 33 per cent of the market while the agricultural banks had another 28 per cent; the two largest commercial banks, ABN and AMRO, each had 10 per cent.[37] As part of this move to joint giro clearing exchange, all the participating banks adopted common transfer forms and a common account numbering system. Even though the commercial banking sector was well organized, it took several years of agonizing discussions before a decision was made. It then took the joint savings banks (who held 15 per cent of the market) another two years before they decided to join this system. Fortunately, the banks started their clearing system before the bulk of growth in retail customers took place. Had they waited longer, the per customer costs would have been significantly higher, as they would have had to switch large amounts of existing customers (and personnel) to the new system. Finally, the existence of a Postgiro system helped by putting competitive pressure on the banks and providing a close to home example of how a giro system could work.

Thus, public action established the first giro system (in the early twentieth century), and a concerted effort by players representing more than half the market established a competing bank giro system the 1960s.

5.2 The economics of giro systems

Exact data on the giro systems are hard to obtain, since the majority of the costs are made by banks, where they tend to be mixed up with other costs. I therefore use data on the Dutch Postgiro from the 1920s, and 1950/60s. This has a double advantage: it provides data on a player whose only business was maintaining accounts for giro clearing, and these data cover the period when the relevant choices were made. Figure 14.3 shows the cost of a giro transfer during these years.

Some interesting observations can be made. There were significant economies of scale in the early years: costs per transaction declined rapidly as volume grew. After a failed attempt to switch to Hollerith technology in 1923, the system remained essentially manual, until 1962.[38] This manual system was surprisingly cheap: in the period 1928–34 it performed on average 52 million transactions per

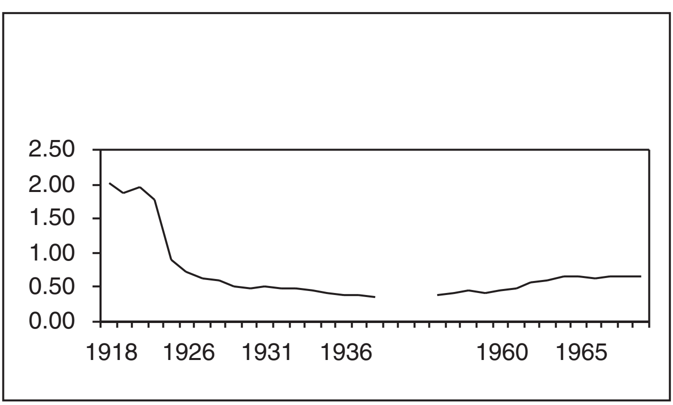


Figure 14.3 Cost of a giro transaction (constant year 2000 € prices).
Source: PCGD annual reports.

year, at an average cost of €26 million per year (converted to current value), or €0.50 per transaction, *much less* than the €0.92 that a mail giro cost in 1994 in Norway (data from Flatraaker and Robinson, 1995, converted to € year 2000 prices).[39] Remarkably, the large-scale automation (1962–4) did not lower per transaction costs; it did however stop the increase caused by the rise in real wages after the war.

Regression analysis using these data suggests fixed costs of running the system of €3–4 million per year, and a per account cost of €110 per year per account. The costs do not appear to significantly depend on the number of transactions.[40] This last observation is important, since this is exactly the cost structure assumed by the model in previous sections of this paper. The model does not assume fixed costs; however the fixed costs in the Dutch giro system were relatively small – they represented about 10 per cent of the total costs in 1936, when the giro had almost 300,000 accounts. Of course as the number of accounts grew, the relative importance of fixed costs declined even further.

5.3 Applying the model

Using these data, let's now examine the decision by Dutch banks to establish their own clearing house in 1967.[41] In particular I try to apply Proposition 1 of the previous section, which stated that a group of banks will only unilaterally adopt a new network technology if $s > c/b$, where s is the joint share of these banks, c is the fixed per customer cost of allowing him to use the new transaction technology and b is the benefit to the bank if a customer uses the new technology for *all* his transactions. Using 1966 data from The Netherlands I estimate c and b to derive the minimum share that is needed to profitably adopt the technology.

In 1966 (the year the decision for a bank giro system was made) total costs of the Postgiro were €254 million, for which they performed 393 million transactions on 1.3 million accounts (these and subsequent figures have all been converted to 2000 price levels). This implies 309 transactions per account and an integral cost of €195 per account. Based on the earlier regression results, I assume that all costs are driven by the number of accounts and not by the number of transactions. Also, given the earlier modest estimates of fixed costs (€3.5 million, or 10 per cent of 1936 cost, and 1.5 per cent of 1966 costs) I assume these fixed costs to be negligible for the purpose of this exercise. Finally I assume that the total costs of the Postgiro system reflect *additional* cost on top of a cheque system.

Thus I get c = €195, where c is the fixed cost per customer per year of the new technology. I now also need an estimate for the extra benefit per transaction of using the new technology, parameter b of the model. To get the value of b, I put the cost of a cheque in 1966 at €1.42.[42] Now $b = (1.42 - 0) \times 309 = 439$ (multiplication with the number of transactions per year is necessary since the model normalized the transactions per person to one). Using Proposition 1 of the previous section, the critical share is then equal to $c/b = 195/439 = 44$ per cent. Thus one or more players need a joint market share of at least 44 per cent to be able to unilaterally and profitably adopt the new giro technology.

The most critical assumption underlying these estimates is of course the cost structure, and in particular the fact that all giro costs are driven by number of accounts, while none of the cheque costs are. To test for the robustness of the estimates against changes in these assumptions, Table 14.1 gives the results for two alternative sets of assumptions. If only 50 per cent of all giro costs are driven by number of accounts (and the remainder of giro costs is transaction related) the critical share decreases to 29 per cent. If only 75 per cent of all cheque costs are transaction driven, the critical market share increases to 59 per cent (if combined, the two alternative assumptions cancel out: the critical share is then 43 per cent).

Comparing these numbers to the market shares of the Dutch banks at the time of the decision to adopt a common giro clearing, we can conclude that the commercial banks (who had only 33 per cent of the market in 1966) did indeed need all of their members and the cooperation of the agricultural banks (28 per cent) to make the economics work. No individual bank had a market share even close to what was needed to go it alone.[43] Finally the hesitancy of the savings banks in joining the initiative (for a while they negotiated to join the Postgiro instead) shows that in the real world other players need not automatically follow, even if a bank or group of banks has decided to adopt a new technology.

In summary, the model appears to do a plausible job of explaining the emergence of giro clearing in The Netherlands. More generally, most countries that adopted giro systems in the early years of the twentieth century are now ACH/giro – rather than cheque countries. Where giro systems did not exist (USA, Canada and UK) banks stuck with the cheque system; the one exception to the rule is France, which had an active giro system in the 1950s and 1960s, but currently still uses cheques on a large scale. Of course, closer analysis of these other countries is needed before we can apply the 'Dutch explanation' elsewhere.

Rather than offer a deterministic explanation for country differences, I would like to make a more general point: analysis of the Dutch events demonstrates that the adoption of payment systems can indeed be characterized by the two properties of network externalities: path-dependence and excess inertia.

Table 14.1 Critical mass for giro adoption in The Netherlands: alternative assumptions

	100% of giro costs and 0% of cheque costs are account driven	*Alternative 1: 50% of giro costs are account driven*	*Alternative 2: 25% of cheque costs are account driven*
Fixed cost per account (€)	195	98	195
Cost per girotransaction (€)	0	0.32	0.00
Cost per cheque (€)	1.42	1.42	1.07
Transactions per person per year	309	309	309
parameter *c*	195	98	195
parameter *b*	439	340	330
Critical share: *c/b*	44%	29%	59%

6 Application to payment harmonization

Let's now turn to the second question raised in Section 1: are differences likely to persist? In this section we split this question in two: (1) are the four cheque countries, and in particular the USA, likely to move from a cheque system to a giro system, and (2) are the European giro countries likely to adopt a common seamless system?

In terms of our model, the first question asks whether a particular country will move from the old technology *f*, to the new technology *g*. The second question asks whether a set of countries that have already adopted less than fully compatible versions of *g* are likely to adopt a common version. According to the model described in Section 4 the answers will to a large extent depend on the level of *autarky* of the countries involved, and on the industry structure within these countries. Below I try to quantify both, and then use the results to answer the two questions.

Autarky. In a world where all consumers transact randomly with each other, for each country i, the share of international payments q_i (as a percentage of all payment transactions in that country) would be equal to $1 - s_i$, where s_i is the share of that country in the world. Thus if Germany represents a fifth of Europe then foreign payments would be 80 per cent of all German payments.[44] As Table 14.2 shows, the share of foreign payments is in reality much lower. Note that in line with our model, the share is higher for small countries, and lower for larger countries. Also, the share is much higher for card payments than it is for transfers. Using the formula $q_i = \delta(1 - s_i)$, where q_i is the share of foreign payments, I get an average δ of 1.6 per cent for the EU (for cards δ is 3.2 per cent, while for transfers alone it is 0.6 per cent). Since $\delta = 1$ corresponds to no autarky and $\delta = 0$ represents total autarky, the results imply that the payments world is still very much a national (or even regional) affair. I lack figures on foreign card payments for

Table 14.2 Share of international transactions for selected EU countries

	Cards	*Transfers*	*Overall*
Belgium	6.5%	0.8%	2.7%
Denmark	2.5%	2.0%	2.4%
Gemany	1.8%	0.2%	0.9%
Spain	1.3%	0.6%	1.1%
France	1.9%	0.1%	0.7%
Ireland	8.6%	0.2%	2.3%
Italy	4.1%	0.4%	1.2%
Netherlands	3.8%	0.5%	2.2%
Austria	5.1%	0.6%	2.2%
Sweden	6.6%	0.5%	2.1%
UK	2.3%	0.2%	1.2%
Average	2.4%	0.3%	1.2%

Source: BIS, EFIMA.[45]

Canada and the US, however for transfers the share of foreign transactions is 0.05 per cent (USA) and 0.2 per cent (Canada) suggesting similar or even larger autarky for these countries.

Industry structure. In most continental European countries banks tend to cooperate in matters of payment networks (this may in part reflect continental Europe's greater tolerance for industry cooperation). Therefore the relevant shares are not those of individual banks, but of all banks cooperating in a network. Within a country, banks typically run one or two networks for ATMs and or debit cards (see Table 14B.2 in Appendix 14B for an overview). Hence for practical purposes I assume that the share of the largest player is in the range 50 to 100 per cent for all countries except the USA, where it is much lower.

Now let's return to the two questions. As was deducted in Section 5, adoption of a giro system requires cooperation of about half the market (44 per cent). Given the low level of international transactions, international considerations are not likely to play a role in the decision to adopt a new technology.[46] This leaves the hurdle at a necessary share of 40–50 per cent. This seems like a high hurdle for the USA, but double for the other three cheque countries, that each have banks organized into one or two payment networks. Accordingly, I would expect US cheques to be around for quite a while.[47]

Turning now to the second question: will Europe adopt a single compatible giro and card system? At present these systems are only partially compatible. In the case of transfers each country has its own system of account numbers, transfer forms, timelines, etc. Similar differences exist in (debit) cards: PIN- *vs*. signature-based systems, different fee structures, disparate data on the magnetic strip, etc. If we assume that in each of the European countries banks decide jointly whether or not to switch to a common standard, we can use the model of Section 4.[48] The players are now countries instead of banks, each with their own 'market share' in the total population. In that case we can apply Proposition 4b, which says that a low δ will create a high probability of lock-in.

In essence countries face a trade-off between costly conversion of international transactions and the cost of switching to a completely uniform system.[49] The joint banks in a country will move to a common system if the per customer costs of adopting this common system are lower than converting all of a customer's international transactions between semi-compatible systems. Below we estimate the current per customer cost of international transactions to get an idea how low the cost of adopting a common system would have to be.

Cost of international transfers. The cost of converting an international transaction between country systems is substantial. Prices for retail cross-border transfers in the Euro zone currently start at €5–7, and banks have indicated that they are willing to go as low as €3. This, as well as anecdotal evidence, suggests that the true costs per retail transaction are indeed in the €5 range. It would cost banks about €3 per customer per year to convert all foreign transfers, even if they would not be able to charge for it.[50]

Cost of international card payments. For card payments there are no detailed data, but Dutch banks have introduced a charge for cross-border ATM usage of

€2–3; again, anecdotal evidence suggests that this is not very different from the transaction costs.[51] It would cost banks €2 per customer per year to convert all foreign card payments.[52]

While these figures represent real money, they may be much smaller than the cost of migrating to a common system (harmonization of account numbers, standardized forms, etc.), even if these costs are spread out over a number of years.

In summary, because the share of international transactions is so low, and conversion costs are relatively modest, the model would predict technological divergence, even if this means conversion of international transactions between incompatible (or semi-compatible) domestic systems. This has indeed been the case for technologies such as cards (where France adopted their chip-based *Carte Bleue*, the UK went with their own credit card scheme called Access, etc.), electronic purse and mobile payments.

7 Conclusions

Two questions were posed in the introduction of this paper: what causes the differences in payment systems and are they likely to persist?

Based on the previous analysis, I would argue that these differences are based on historical coincidences such as the emergence of giro systems in the 1920s. The network nature of many payment systems has caused these differences to persist and at times increase over the past century. Looking forward, these differences are likely to persist rather than disappear in the short or even medium term.

Overall these findings have some important implications:

1 Cooperation between players may prevent lock-in into inferior technologies. Consequently, a certain tolerance for such cooperation may be asked from the regulator.
2 A pan-European payment network may be further off than most people think. Instead, a continuing patchwork using converters is more likely. This means that cross-border payments will likely continue to cost (much) more than national payments.
3 If banks or regulators really want pan-European solutions, they will need strong European cooperation platforms to offset the existing national payment organizations.
4 Outside players that want to enter the European payments arena (credit card issuers, Internet-payment players like Paypal) may be well advised to take a country-by-country approach. In spite of a common European currency, for the time being ‘one size *will not* fit all’.

As a concluding remark, this paper offers only the beginning of an explanation of the differences in payment technology. Interesting questions remain to be answered by future work, for example with regard to card technologies: why did most of the cheque countries adopt credit cards (including the USA, in spite of –

or thanks to? – a fragmented payment system), where continental Europe adopted debit cards? Will these technologies converge, will all countries adopt a mix of debit and credit cards, or will even the USA move from credit to debit cards?

Appendix 14A: Proof of Propositions 1–4

> *Proposition 1. 'Fragmentation increases the risk of lock-in': A welfare sub-optimal Nash-equilibrium exists if and only if the market share of the largest player is lower than the cost/benefit ratio of the new technology* g.

Proof: Each bank i faces the decision whether or not to adopt technology g, by maximizing its utility. Since the amount of customers per bank is fixed, this is equivalent to maximizing its utility per customer: $Ui\ (s_i, s_g, t_i)$. This utility is a function of three parameters: (1) the market share of bank i, denoted s_i; (2) the combined share s_g of other banks using g; and (3) t_i, which represents the technology used by bank i, where $t = g$ if the bank adopts g and $t = f$ otherwise. Now:

$$U_i(s_i, s_g, f) = 0 \qquad \text{if the bank does not adopt } g$$
$$U_i(s_i, s_g, g) = (s_i + s_g)b - c \qquad \text{if it does adopt } g.$$

Since by definition $b > c$, $s_g = 1$ (all banks adopt g) is always a Nash-equilibrium: unilateral deviation (dropping or not adopting g) from this equilibrium will lower a bank's utility from a positive number to zero. Now let s_1 be the market share of the largest bank. Then there is a second equilibrium where $s_g = 0$ (no bank adopts g) if:

$$s_1 b - c < 0 \quad \Leftrightarrow \quad s_1 < c/b \tag{1}$$

This is a Nash-equilibrium because if (1) holds for the largest player it automatically holds for all other players. This equilibrium is welfare suboptimal: if all banks adopt g, they all have positive per customer benefits of $b - c > 0$; if no bank adopts g, each bank has per customer benefits of zero. Finally, $s_g = 0$ and $s_g = 1$ represent the *only* Nash-equilibriums. To see this, consider the situation where $0 < s_g < 1$, i.e. some but not all banks have adopted g. Then if $s_g < c/b$ the benefits to each bank that uses g are $s_g b - c < 0$, so each one of these banks would be better off by dropping g. However if $s_g \geq c/b$ then each bank i that does not use g would have adoption benefits of $(s_g + s_i)b - c > s_g b - c \geq 0$, because $s_g \geq c/b$. Thus $0 < s_g < 1$ cannot be a Nash-equilibrium.

> *Proposition 2. 'Upgrades increase lock-in': The availability of upgrades increases the number of situations where welfare suboptimal equilibriums occur.*

Proof: Let F be an upgraded version of f. A bank can switch its customers to technology F at a fixed cost per customer c_F and can then get benefits per transaction of b_F. Assume $c_F < c$ and $b_F - c_F < b - c$ (and therefore $b_F < b$): F can be obtained at a lower cost, but the ultimate benefit is lower than the potential benefit of g. However a crucial difference is that F is compatible with f, so it can be used for all transactions, where g can only be used if both parties have adopted g.

When upgrading is an option, the per customer utility for bank i becomes:

$$
\begin{array}{lll}
U_i(s_i, s_g, f) & = 0 & \text{if it adopts neither } F \text{ nor } g \\
U_i(s_i, s_g, F) & = b_F - c_F & \text{if it adopts } F \\
U_i(s_i, s_g, g) & = (s_i + s_g)b - c & \text{if it adopts } g \\
U_i(s_i, s_g, Fg) & = b_F + (s_i + s_g)(b - b_F) - c - c_F & \text{if it adopts both } F \text{ and } g.
\end{array}
$$

Note that $U_i(s_i, s_g, F) > U_i(s_i, s_g, f)$ so that any bank will always prefer upgrading to F over sticking with f. Also note that $U_i(s_i, s_g, F) < U_i(s_i, 1, g)$ so the situation where all banks upgrade is a welfare suboptimal. However, all banks upgrading to F is a (suboptimal) equilibrium if $U_i(s_i, 0, F) \geq U_i(s_i, 0, g)$ and $U_i(s_i, 0, F) \geq U_i(s_i, 0, Fg)$ for all i. Or:

$$b_F - c_F \;\geq\; s_i\, b - c, \;\; \text{for all } i \Leftrightarrow s_1 \;\leq\; c/b + (b_F - c_F)/b \tag{2}$$

and:

$$b_F - c_F \;\geq\; b_F + s_i(b - b_F) - (c_F + c), \;\; \text{for all } i \Leftrightarrow s_1 \;\leq\; c/(b - b_F) \tag{3}$$

where s_1 is the share of the largest bank. Now the right-hand side of both (2) and (3) is larger than that of (1), so there is now a larger area where a suboptimal equilibrium occurs.

Figure 14A.1 illustrates how upgrading can increase the area where lock-in occurs.[53] The horizontal axis is s_1 the market share of the largest bank, while the vertical axis depicts the cost/benefit ratio of g. Without upgrades lock-in can occur only in the upper-left triangle. The availability of upgrades enlarges this area with the flat triangle below the diagonal.

> *Proposition 3. 'Autarky promotes innovation': A low δ will decrease the occurrence of lock-in into an inferior technology.*

Proof: If the total share of other banks that use g is s_g, then any bank i adopting g can use it for a fraction of its transactions equal to:

$$(1 - q_i) + q_i s_g/(1 - s_i)$$

The first term represents the inhouse transactions, the second the interbank transactions. After substituting $q_i = \delta(1 - s_i)$ and taking $s_g = 0$ we get the utility of adopting g for the largest player if nobody else uses g:

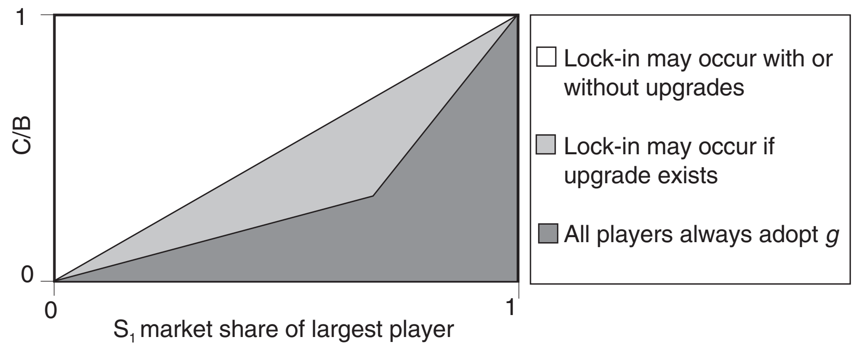


Figure 14A.1 Occurrence of lock-in with upgrades.

$$U_1\ (s_1, 0, g) = [1 - \delta(1 - s_1)]\ b - c \tag{4}$$

Now lock-in requires that $U_1(s_1, 0, g) < 0$, or $1 - \delta(1 - s_1) < c/b$. But $1 - \delta(1 - s_1) > s_1$ because $1 - \delta(1 - s_1) = s_1 + (1 - \delta)(1 - s_1)$. Thus a lower δ decreases the area where non-adoption is a Nash-equilibrium.[54]

Proposition 4. 'Autarky breeds diversity': (a) Multiple incompatible versions of a network technology g *can form a Nash-equilibrium if and only if* $\delta < 1$ *and there are positive costs to switching between versions of* g*; (b) the probability of the adoption of incompatible versions increases as* δ *decreases; (c) if autarkic players* ($\delta < 1$) *adopt incompatible versions, and* δ *subsequently rises to 1, the players will always pass through a phase where no individual player will migrate to a common technology, while a switch to the standard of the largest player would enhance overall welfare.*

Proof: Assume the new technology comes in several incompatible versions: $g_1, g_2, \ldots, g_n$, all with similar costs and benefits. Also assume that once a bank has adopted version g_i, it can switch to another version g_j at an extra cost per customer of c_m. Assume $c_m \leq c$.[55] If two banks have incompatible versions of g, they have to revert back to the old technology f for these transactions. For each player the utility function now becomes:

$U_i\ (s_i, s_g, f)$	$= 0$	it sticks with f
$U_i\ (s_i, s_g, g_j)$	$= [1 - q_i + q_i\, s_g / (1 - s_i)]b - c$	it *adopts* any version g_j
$U_i\ (s_i, s_g, g_j)$	$= [1 - q_i + q_i\, s_g / (1 - s_i)]b - c - c_m$	it *switches* to another version g_j

We again look for the occurrence of suboptimal Nash-equilibriums. Obviously, the optimal Nash-equilibrium is $s_g = 1$ and $t_i = g$ for any specific version of g: all players adopt the same version of g. Following Proposition 3, $t_i = f$ for all i (all banks stick with f) is a (suboptimal) equilibrium if $c/b > 1 - \delta(1 - s_1)$, where s_1 is the share of the largest bank. However, now there may also be equilibriums where banks adopt different versions of g. For example each bank i adopting and maintaining its own version g_i is an equilibrium if:

$$(1 - q_i)b - c \; > \; [(1 - q_i) + q_i s_1/(1 - s_i)]b - c - c_m \qquad \text{for all } i \neq 1 \tag{5}$$

and:

$$(1 - q_i)b - c \; > \; 0 \qquad \text{for all } i \tag{6}$$

(5) requires that for each player i maintaining its own version g_i is more attractive than switching to g_1 (the standard of the largest player); (6) requires that for each bank adoption of its own version, g_i is more attractive than sticking with f. Substituting $q_i = \delta(1 - s_i)$ and rearranging terms we get:

$$\delta s_1 \; < \; c_m/b \text{ and } 1 - \delta(1 - s_i) \; > \; c/b \tag{7}$$

Note that if $\delta = 1$, (7) becomes $s_1 < c_m/b$ and $s_i > c/b$, which has no solution since $c \geq c_m$ and $s_1 \geq s_i$. Thus incompatible versions can only occur if players are to some extent autarkic. However if $\delta < 1$, there always exist parameter combinations that meet (7). For example if $\delta = 0.1$ any combination of player sizes enables an equilibrium where each player has its own standard, as long as $0.1 < c_m/b \leq c/b < 0.9$. This proves part (a) of the proposition.

According to (7) a lower δ will decrease the lower bound for c_m/b and increase the upper bound for c/b. This proves part (b) of Proposition 4.

Turning now to part (c), the welfare effects. If all players stick to their own version g_i, overall per customer utility across all banks is equal to the difference between these expressions:

$$\sum_i s_i \, U_i \,(s_i, s_i, g_i) \; = \; \sum s_i \,(1 - q_i)\, b - c \; = \; [1 - \delta\,(1 - \mathrm{H})]b - c$$

$\mathrm{H} = \sum_i s_i^2$ is the Herfindahl index. If on the other hand all players adopt the version of the largest player, overall utility is equal to

$$\sum s_i \, U_i \,(s_i, 1, g_1) \; = \; b - c - (1 - s_1)\, c_m$$

Thus the net welfare gain of switching to g_1 is equal to:

$$\delta\,(1 - \mathrm{H})\, b - (1 - s_1)\, c_m.$$

This is positive if:

$$c_m/b \; < \; \delta\,(1 - \mathrm{H})/(1 - s_1) \tag{8}$$

Now if s_1 is the share of the largest player, H can have values from $s_1^{\,2}$ to $s_1^{\,2} + (1 - s_1)^2$, depending on the industry structure. So the lowest value of the right-hand side of (8) is $2\delta s_1$; this is obtained by replacing H in (8) with its upper boundary: $s_1^{\,2} + (1 - s_1)^2$. From (7) we know that different versions is an equilibrium if $c_m/b > \delta s_1$. Thus there is always a positive lock-in region $\delta s_1 < c_m/b < 2\delta s_1$ where nobody switches, but overall welfare would enhance if all would adopt the standard of the largest player.

Appendix 14B: Multiple players in semi-connected countries

In this section I explore the situation where there are multiple countries subject to some autarky ($\delta < 1$ *across* countries), and several banks within each country. I assume that $\delta = 1$ *within* countries.

As mentioned earlier, different versions of g *within* a country cannot be an equilibrium if $\delta = 1$. I therefore look for the occurrence of two other non-optimal equilibriums: (1) all banks in all countries stick to f, and (2) countries adopt different versions of g (but all banks in a given country adopt the same version of g).

Propositions 5 and 6 confirm that the essence of Propositions 3 and 4 holds in this more complex environment, with an added role for the industry structure within each country.

> *Proposition 5. 'Organizing banks into a set of somewhat autarkic regions makes the adoption of a new network technology more likely': The number of situations where lock-in into* f *is an equilibrium, decreases with: a lower* δ*, higher concentration of countries and higher concentration of banks within countries.*

Proof: Let there be M countries, with share s_i, $i = 1 \ldots M$, and within each country N banks with national share t_{ij}, $j = 1 \ldots$ N, where $\sum_j t_{ij} = 1$ for all i, and $\sum_{ij} s_i t_{ij} = 1$. Now non-adoption of g is an equilibrium if for all i, j:

$$(1 - q_i)t_{ij}b - c \;<\; 0 \;\Leftrightarrow\; c/b \;>\; [1 - \delta(1 - s_i)]t_{ij} \tag{9}$$

Note that the right-hand side of the second inequality increases with s_i (country concentration) and t_{ij} (industry concentration in a country) and decreases with d.

If these same players would exist without a country superstructure being imposed, that is $\delta = 1$, bank ij would have an overall market share $s_i t_{ij}$ and non-adoption would be an equilibrium if for all i, j:

$$s_i t_{ij} b - c \;<\; 0 \;\Leftrightarrow\; c/b \;>\; s_i t_{ij} \tag{10}$$

Now the last term in (10) is smaller than the last term in (9) if $s_i t_{ij} < [1 - \delta(1 - s_i)]t_{ij}$, which is always the case if $\delta < 1$ and $s_i < 1$. This means that there are values of c/b for which banks in a country structure would adopt g, while without such a structure banks may be locked in to f. The difference between the two expressions gives an indication of the magnitude of the influence of a country superstructure:

$$[1-\delta(1-s_i)]t_{ij}-s_i\,t_{ij} \;=\; (1-\delta)(1-s_i)t_{ij}$$

Even for modest parameter values this influence can be large. For example suppose $\delta = 0.2$, $s_i = 0.1$ and $t_{ij} = 0.5$ (ten countries each with two players of equal size); the difference is then equal to 0.36: a cost–benefit ratio of 0.41 is sufficient to move the countries to *g*, while a ratio of 0.05 would be needed to get the banks to adopt *g* in the absence of a country structure.

These same factors also foster the adoption of incompatible versions of *g*.

> *Proposition 6. The number of situations where lock-in into incompatible versions of* g *is an equilibrium, increases with: a lower* δ, *higher concentration of banks within countries and smaller difference in size between the largest and the smallest country.*

Proof: Assume the share of the market leader in each country is t_1. We are looking for a situation where the leader in at least one market adopts some version of *g*, thus by proposition 5, expression (9):

$$c/b \;<\; [1-\delta(1-s_i)]t_1 \qquad \text{for at least one } i \tag{11}$$

Without loss of generality let this be the largest country, with share s_1. In addition we want no (market leading) bank to have an incentive to switch to the standard of the leading country, so by rewriting (5):

$$\begin{aligned}&(1-q_i)\,b-c \;\geq\; [(1-q_i)\,t_1+q_i\,s_1/(1-s_i)]\,b-c-c_m \;\Leftrightarrow\\ &c_m/b > \delta\,s_1-(1-t_1)\,[1-\delta(1-s_i)] \qquad \text{for all } i \neq 1, j\end{aligned} \tag{12}$$

By combining (11) and (12), we get the following sufficient condition for the existence of such lock-in:

$$[1-\delta(1-s_1)]\;t_1 > c/b \geq c_m/b > \delta s_1-(1-t_1)\;[1-\delta(1-s_i)] \;\text{ for all } i \neq 1, j \tag{13}$$

So as long as the left-most term in this expression is smaller than the right-most term, values of $c/b \geq c_m/b$ exist for which lock-in can occur. After some manipulation (13) can be written as:

$$(1-t_1)(s_1-s_i) \;<\; 1/\delta-1 \tag{14}$$

As country 1 is the largest country, $s_1 - s_i < 1$. Thus, a sufficient condition for (14) to hold is $t_1 > 2 - 1/\delta$. The values of number of c_m/b for which lock-in can occur increases as the two sides of (14) grow further apart, that is the occurrence of lock-in increases with: t_1 (industry concentration in each country), and with $(s_1 - s_i)$: if all countries are of equal size lock-in is more likely. Finally, as before, lock-in becomes more likely if δ is low.

The intuition behind these last two proposals is fairly straightforward, given Propositions 3 and 4. The role of country concentration (measured by s_1) merits some comment. For the move from f to g (Proposition 5), what counts are absolute shares of the countries. What counts for the move to a common standard however (Proposition 6), is the difference between the largest and the smallest player; a single large country provides a natural standard for others, while an oligopoly of equal sized countries can result in deadlock.

Tables on giro systems and number of networks

Table 14A.1 Giro systems in Europe

Country	*Year of introduction*	*Giro transactions (millions, 1958)*	*Transactions per capita (1958)*
Austria	1883	131	18
Belgium	1913	47	5
Denmark	1920	76	17
Finland	1940	33	7
France	1918	706	14
Germany	1908	1,027	14
Netherlands	1918	295	25
Italy	1918	n/a	n/a
Luxembourg	1911	n/a	n/a
Norway	1942	37	10
Sweden	1925	218	29
Switzerland	1906	253	45
Total		2,823	16

Source: Thompson (1964)

Table 14A.2 Number of payment networks[56]

	1988	*1999*
Belgium	4	2
Canada	4	n/a
France	1	1
Germany	4	4
Italy	1	1
Netherlands	2	1
UK	3	1
EU Average	n/a	2
US	37	23

Source: BIS, ECB.

Notes

1 I have corrected the BIS data for Germany to reflect the fact that about half the German debit card transactions are recorded as direct debits by BIS.

2 According to the most recent BIS estimates the number of cheques *per capita* in the USA went from 244 in 1998 to 245 in 1999 (BIS, 2001). See Murphy (1991) for the quote on cheques used to pay credit card bills.

3 Figures from Wells (1996). Flatraaker and Robinson (1995) report average costs (at the bank level) for Norway: €1.75 per cheque *vs*. €0.81 for a giro transfer.

4 Humphrey, Pulley and Vesala (2000).

5 Humphrey, Kim and Vale (2000).

6 Humphrey and Berger (1990).

7 Evans and Schmalensee (1999).

8 For example, the Dutch competition authority (NMa) recently argued that banks should offer their customers the ability to take their account number with them to another bank (NEI, 2000).

9 Baumol (1952). For a comprehensive survey, see Boeschoten (1992). More recent contributions include Santomero and Seater (1996), and Attanasio, Guiso and Japelli (1998).

10 See for example McAndrews and Roberds (1999), who conclude that cheque float acts as a distorting tax, if it is not reflected in merchant prices. Shy and Tarkka (1998) model the trade-off between credit cards and smart cards.

11 The authors regress annual *per capita* transactions for each instrument on explanatory variables like the price of the various instruments, *per capita* income, number of ATMs and debit card terminals, currency holdings per person, crime rate and bank-sector concentration. While they find significant relationships with variables like number of ATMs and debit card terminals, they acknowledge that these explanatory variables are largely endogenous. On the other hand the explanatory value of truly exogenous variables like crime rate and income is quite modest.

12 Real dollar terms: figures are corrected for inflation. The decline is due to lower interest rates and quicker processing.

13 Murphy (1991) estimates the impact of per item charges on cheque usage and finds it minimal: correcting for factors like age, income, etc., per item charges reduce cheque usage by 10 per cent. This suggests that perverse pricing can at best explain a small part of cheque usage. Humphrey, Kim and Vale (2000) find a strong correlation between

cheque decline and strong price incentives to use cards. However, since other European countries showed the same decline in cheques over that period without Norwegian price differences, the causality may be weaker than the authors conclude.

14 Figure from Bauer and Ferrier (1996).

15 See Bauer and Ferrier (1996) for both findings.

16 Speech held at the National Automated Clearing House Association (NACHA) 2000 annual meeting.

17 Boeschoten and Fase (1989), Mot, Cramer and van der Gulik (1989). In terms of number of transactions, 'cash is (still) king': the average person makes over 1,000 payments a year, of which over 90 per cent have a value below €5 (Boeschoten, 1992).

18 If a consumer adopts a standard, he increases the value of the standard for all other users. If the user is not financially compensated for this, it creates a positive externality. This effect is analogous to the better known negative externalities such as road congestion, air pollution, etc.

19 The seminal papers are Katz and Shapiro (1985), Farrell and Saloner (1986) and Arthur (1989). They all analyse goods with consumer benefits of the form $a + b(n)$, with a being the stand-alone value of the good and $b(n)$ the benefit of the network, depending of the number of other users of that same (or compatible) good.

20 Quote from Matutes and Regibeau (1996), who give a comprehensive overview.

21 While obvious to most non-economists, the relevance of this phenomenon is disputed within the profession. Opponents argue that path dependence is either much less pervasive than proponents suggest, or irrelevant because it does not help understand and analyse economic reality, essentially converting economics into the study of economic history (see for example Liebowitz and Margolis, 1994).

22 See Economides and Himmeberg (1995) for the US fax market, Dranove and Gandal (1999) for DVDs, Park (2000) for VCRs, and Saloner and Shepard (1995) for ATMs. The ATM application analyses the adoption of ATMs by banks, and finds that US banks with a larger branch network (i.e. more than five branches!) have adopted ATMs faster, supposedly because they could install a larger ATM network (one in each branch) giving a higher benefit to their customers.

23 Cowan (1990) analyses nuclear technology, and finds that due to historical coincidences (use of nuclear reactors on submarines) the US settled on what is now considered an inferior technology. See David (1985) for the QWERTY case, and Liebowitz and Margolis (1994) for a rebuttal.

24 Fase (1999) and Shy (2001).

25 The interchange fee is charged between banks that participate in a network transaction. For example in credit cards, the bank of the card-accepting merchant has to pay 1 to 2 per cent of the transaction value to the bank of the customer that used his card. See Matutes and Padilla (1994) for a model that analyses the decision of banks to link their ATM networks. A similar approach is explored for other channels by Bouckaert and Degryse (1995) for phone banking and Vesala (2000) for remote channels in general.

26 Bassanini and Dosi (1998) and Cowan and Cowan (1998).

27 I normalize these costs to reflect the fixed cost per customer per unit of time during which the average customer makes one transaction.

28 Alternatively they could represent increased customer fees made possible because of the higher customer value of the new technology g.

29 Numbers from the Dutch banking system confirm that this assumption holds for the Dutch market.

30 Assumption 1 is introduced purely for simplicity; it is not required to derive the results. Assumption 2 is in line with the assumption of constant hedonic prices (see next note).

31 Alternatively, assume that any increase/decrease in customer benefits due to the new technology is compensated through an increase/decrease in price (constant hedonic prices).

32 More precisely, they can at best re-establish the 'pre-*g*' profit level by both adopting compatible versions of *g* (which implies that all the net benefits are passed onto the consumer). The only exception is if the network effect *b* is so large compared to the 'distance' (differences in product offering or switching cost) between firms that one bank can use *g* to drive the other out of business (which is not observed in practice). The underlying model and proof is available from the author upon request.

33 Throughout the remainder of this section, I will define welfare as the joint welfare of all banks. The model takes consumer welfare to be unaffected by the choice of technology, by assuming that any impact on consumer welfare is compensated by the price charged to the consumer. And since this price is in turn reflected in b, changes in consumer welfare are 'absorbed' by banks.

34 See C.K. Harley (1970) for the sailing ship effect. In the payment system practice the lock-in effect is further enhanced by the fact that two systems have to be maintained for *f* and *g*, each with their own fixed costs; for simplicity our model assumes these fixed costs (other than the fixed cost per customer for *g*) to be negligible. To the extent that they are non-negligible they will enhance the lock-in effect.

35 The notable exceptions are the UK and its former colonies (US and the commonwealth). See Table 14A.1 in Appendix 14B for an overview based on Thompson (1964).

36 The two systems remained separate until the (late) 1990s, when following increasing pressure from the Dutch National Bank they were finally made fully compatible. In 1966 the Postgiro system had a share of about 23 per cent (this includes PCGD, GGA and RPS).

37 Figures reflect share of commercial and agricultural banks of non-interbank deposits as of December 1966 (figures from De Nederlandsche Bank). In 1964 two mega-mergers of four players with each about 5 per cent share created ABN and AMRO who then had about 10 per cent each. This merger did likely facilitate the decision to make a major move into the mass market and set up an interbank giro clearing system.

38 The story of this failed switch is quite colourful. The failure was so complete, that the Postgiro had to shut down operations for a full year to sort out the mess. It even had to ask customers to provide proof of their account balances, because the internal administration could no longer be relied upon. An official inquiry found many causes, among them an overly optimistic external consultant, who had a contract entitling him to a quarter of all labour savings during the first year (Tak and Dubois, 1924).

39 The Postgiro transaction figures include cash withdrawals, deposits and a few cheques, but the majority of transactions were giro transfers. Several explanations can be offered for this seemingly paradoxical result. First real labour costs more than quadrupled in the period 1928–94. Second, the functionality of a 1994 giro payment was much richer than in 1928: printed statements, added information, etc.

40 Regressing data on total (real) costs (COST) on the number of accounts (ACC) and transactions (TX) for the period 1918–1928 for the Dutch giro bank, yields:

$$\underset{(t = 2.6)}{\text{COST} = 3.5\text{ million}} + \underset{(t = 3.0)}{110\ \text{ACC}} + \underset{(t = 0.3)}{0.034\ \text{TX}} \qquad \text{Adj. } R^2 = 96\%\ ,\ \text{DW} = 1.24$$

41 Table 14A.1 in the appendix confirms that The Netherlands is a good representative of the more successful giro countries.

42 I get this figure by applying a multiple of 2.2 (based on both the Norwegian and the US experience) on the integral cost of a giro transfer. Using the 1966 giro data I get an *integral* cost of €0.64 per giro transaction (by assuming that the variable cost is zero). Note that I implicitly assume that all the costs of a cheque are transaction rather than account related.

43 In 1918 the Postgiro was established with a much lower share. However, first this was a public initiative, and second it may be argued that this institution was to a large extent autarkic, which significantly reduces the necessary critical market share.

44 Germany represents 22 per cent of EU population. Of course the calculation in the text only holds if Europe would form a closed system.

45 The data on cross-border transfers are based on BIS data for domestic transactions and category 1 Swift messages. Each international transfer is assumed to generate two such messages. Data for international card payments based on EFIMA estimates for 1997. Portugal, Finland, Luxembourg and Greece were omitted because of data problems.
46 Theoretically, this number is influenced by the systems used by other countries: if all countries use the new technology, less than 44 per cent would be needed: 43.8 per cent to be exact if $\delta = 0.6$ per cent and the share of the country is 1/3 (see Appendix B for the detailed mechanics).
47 Checks may receive an additional 'life-extension' through upgrades. The impact of upgrades has so far been limited, as evidenced by the analysis of technology change in cheque processing that was mentioned in Section 2. However a major upgrade may currently be presenting itself: now that most banks retain computer images of cheques, and congress is expected to pass legislation confirming the legal status of a cheque image, it becomes feasible to transfer images rather than paper between banks. This is expected to substantially reduce the cost of cheque processing.
48 If banks do *not* act jointly, we need to use the more complicated model of Appendix B. However, for the case at hand, this does not lead to different conclusions.
49 Of course the fact that the customer bears the conversion costs, while the banks would have to swallow most of the cost of switching to a common system makes this trade-off quite simple. But even if we assume the banks would have to bear the full cost of converting international transactions, adopting a common system would have to be quite cheap, before the banks would do it.
50 Based on an average of 95 transfers, direct debits and cheques per person per year (current EU average), of which 0.3 per cent are cross border, at a cost of €5–7 each.
51 The prices mentioned are for retail transactions in the euro zone where the banks no longer make a Foreign Exchange conversion margin. Traditionally the FX margin on (wholesale) transactions was the major source of income on cross-border transactions.
52 Based on an average of 25 card transactions per person per year (current EU average), of which 5.7 per cent are cross-border, at a cost of €2–3 each.
53 For the sake of the graph it is assumed that both the benefits of the upgrade are half those of *g*.
54 Note that like before there is no Nash-equilibrium where some but not all players adopt *g*. Since the right-hand side of (4) is increasing in s_g, if (4) holds for one player, it will automatically lead to a positive $U_i\ (s_i,\ s_i{+}s_g,\ g)$ for all other players.
55 This last condition says that migrating to another version of *g* can never be more expensive than adopting *g* in the first place, which seems reasonable.
56 The figures in the table are for each country the minimum number of ATM networks and debit card (EFTPOS) networks.

References

Arthur, B.W. (1989) 'Competing technologies, increasing returns, and lock-in by historical events', *Economic Journal* 99: 116–31.

Attanasio, O.P.L. *et al.* (1998) 'The demand for money, financial innovation, and the welfare cost of inflation: an analysis with households' data', NBER Working Paper 6593.

Bassanini, A.P. and Dosi, G. (1998) 'Competing technologies, international diffusion and the rate of convergence to a stable market structure', IIASA Interim Report, no. IR-98-012.

Bauer, P.W. and Ferrier, G.D. (1996) 'Scale economies, cost efficiencies, and technological change infederal reserve payments processing', *Journal of Money, Credit and Banking* 28(4): 1004–39.

Baumol, W. J. (1952) ‘The transactions demand for cash: an inventory theoretic approach’, *Quarterly Journal of Economics* 66: 545–56.

BIS (2001) ‘Statistics on payment systems in the group of 10 countries (red book)’, Bank for International Settlements, Basel.

Boeschoten, W.C. (1992) ‘Currency use and payment patterns’, Doctorate dissertation, University of Amsterdam.

Boeschoten, W.C. and Fase, M.M.G. (1989) ‘The way we pay with money’, *Journal of Business & Economic Statistics*, 1989/7: 319–26.

Bouckaert, J. and Degryse, H. (1995) ‘Phonebanking’, *European Economic Review* 39: 229–44.

Cowan, R. (1990) ‘Nuclear power reactors: a study in technological lock-in’, *Journal of Economic History* 50(3): 541–67.

Cowan, R. and Cowan, W. (1998) ‘Technological standardization with and without borders in an interacting agents model’, Merit Research Memorandum # 2/98–018.

David, P.A. (1985) ‘Clio and the economics of QWERTY’, *Economic History* 75(2): 333–7.

Dranove, D. and Gandal, N. (1999) ‘Network effects, standardization, and the Internet: what have we learned from the DVD vs. DIVX battle?’, CEPR Discussion Paper 2335.

ECB (1996 and 2000) ‘Payment systems in the European union’, (‘blue books’), ECB, Frankfurt.

Economides, N. and Himmelberg, C. (1995) ‘Critical mass and network size with application to the US fax market’, mimeo, New York University, EC-95–11.

Evans, D. and Schmalensee, R. (1999) *Paying with Plastic, the Digital Revolution in Buying and Borrowing*. MIT Press.

Farrell, J. and Saloner, G. (1986) ‘Installed base and compatibility: innovation, product preannouncements, and predation’, *American Economic Review* 76(5): 940–55.

Fase, M.M.F. (1999) *Geld in het fin de siècle*. Amsterdam University Press.

Flatraaker, D. and Robinson, P. (1995) ‘Income, costs and pricing in the payment system’, *Norges Bank Economic Bulletin* 66: 321–32.

Harley, C.K. (1970) ‘British shipbuilding and merchant shipping: 1850–1890’, *Journal of Economic History* 30(1): 262–6.

Humphrey, D.B., Kim, M. and Vale, B. (2000) ‘Realizing the gains from electronic payments: costs, pricing and payment choice’, Working Paper.

Humphrey, D.B. and Berger, A.N. (1990) ‘Market failure and resource use: economic incentives to use different payment mechanisms’, *The US Payment System: Efficiency, Risk and the Role of the Federal Reserve*, pp. 45–86.

Humphrey, D.B., Pulley, L.B. and Vesala, J. (1996) ‘Cash, paper, and electronic payments: a cross-country analysis’, *Journal of Money, Credit and Banking* 28(4): 914–39.

Humphrey, D.B., Pulley, L.B. and Vesala, J. (2000) ‘The check’s in the mail: why the United States lags in the adoption of cost-saving electronic payments’, *Journal of Financial Services Research* 17(1): 17–39.

Katz, M.L. and Shapiro, C. (1985) ‘Network externalities, competition, and compatibility’, *American Economic Review* 75(3): 424–40.

Liebowitz, S.J. and Margolis, S. (1994) Network externality: an uncommon tragedy, *Journal of Economic Perspectives* 8(2): 133–50.

Matutes, C. and Padilla, A.J. (1994) ‘Shared ATM networks and banking competition’, *European Economic Review* 38: 1113–38.

Matutes, C. and Regibeau, P. (1996) ‘A selective overview of the economics of standardization. Entry deterrence, technological progress and international competition’, *European Journal of Political Economy* 12: 183–209.

McAndrews, J.J. and Roberds, W. (1999) ‘A general equilibrium analysis of check float’, Working Paper, Federal Reserve bank of New York.

Mot, E.S., Cramer, J.S. and van der Gulik, E.M. (1989) ‘De Keuze van een betaalmiddel’, Research memorandum, Stichting voor Economisch Onderzoek, UvA.

Murphy, N.B. (1991) ‘Determinants of household check writing: the impacts of the use of electronic banking services and alternative pricing of services’, *Financial Services Services Review* 1(1): 35–44.

NEI Kenniscentrum voor Marktwerking en Mededinging (2000) ‘Fusies en overnames in het Nederlandse bankwezen; finaal rapport’, *De Nederlandse Mededingingsautoriteit*, July 2000.

Park, S. (2000) ‘Quantitative analysis of network externalities in competing technologies: the VCR case’, Mimeo, SUNY at Stony Brook.

PostCheque&GiroDienst, Annual Reports, The Hague.

Saloner, G. and Shepard, A. (1995) ‘Adoption of technologies with network effects: an empirical examination of automated teller machines’, *RAND Journal of Economics*, 26(3): 479–501.

Santomero, A.M. and Seater, J.J. (1996) ‘Alternative monies and the demand for media of exchange’, *Journal of Money, Credit and Banking* 28(4): 942–60.

Shy, O. (2001), *The Economics of Network Industries*. Cambridge University Press.

Shy, O. and Tarkka, J. (1998) ‘The market for electronic cash cards’, Bank of Finland Discussion Papers, vol. 21/98.

Tak, A. and Dubois, F.R.J. (1924) ‘Rapport over de oorzaken van de ontwrichting en de verergerde ontwrichting van den postchèque – en girodienst en de schuldigen daaraan’, The Hague, Algemeene Landsdrukkerij.

Thomson, F.P. (1964) *Giro Credit Transfer Systems*. Oxford: Pergamon Press.

Wells, K.E. (1996) ‘Are checks overused?’, *Federal Reserve Bank of Minneapolis, Quarterly Review* 20(4): 2–12.

15 Electronic payments and scale properties in the banking industry*

Kjersti-Gro Lindquist

Abstract

As a result of new technology, payment services in banking have shifted dramatically from paper-based giro and cheque payments to electronic giro and debit card payments. This paper analyses the importance of new electronic payment systems for the development in banks' scale properties and input demand in their production of loans and deposits. A four-factor translog cost function and the corresponding cost-share equations are estimated. Labour, physical capital, materials and funding are treated as variable inputs. The results show that the increase in electronic payments has increased the elasticity of scale and decreased average variable costs in banking. The move towards electronic payment systems has affected input demand asymmetric, i.e. non-neutral, and the input ratio between labour and both physical capital and materials have decreased.

1 Introduction

Over the last couple of decades, banks in Norway, as in most OECD countries, have gone through important changes. Market deregulation, increased competition, technological innovations and financial crises have spurred mergers and acquisitions – both domestic and across borders, investments in new technologies and probably also changed banks' strategic behaviour. While several of these issues have been analysed and discussed in the literature, little has been done on the importance of new electronic payment systems, which clearly represent new technology.

As a result of this new technology, payment services have shifted dramatically from paper-based giro and cheque payments to electronic giro and debit card payments. While only 10 per cent of non-cash payments in Norway were in electronic form in 1987, by 1999 this had risen to 60 per cent. This change was – at least partly – spurred by banks' pricing policy; see Humphrey *et al*. (2001). Banks' motivation to offer electronic payment systems may be due to both cost saving and competition: once the necessary investment is made, for the same type of transaction, electronic payments cost banks much less to produce than paper-based payments; see Flatraaker and Robinson (1995) and Wells (1996). If customers find

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

electronic payment systems more convenient than paper-based systems, and hence prefer the first, electronic payment systems may be viewed as a strategic variable in the competition for customers.

This paper focuses on the importance of new technologies, i.e. new electronic payment systems, for the development in banks' scale properties and input demand in the production of loans and deposits. As in general in analyses of bank behaviour, the definition of bank output is a major challenge, and one may argue that payment services, including both electronic, paper-based and cash payment services, should be treated as an output. However, payment services contribute marginally to banks' income, and our view is that these services have the character of being inputs in the production of loans and deposits, because banks must offer these services to attract depositors in particular. Payment services are either 'intermediate' inputs produced within banks by the primary inputs labour, physical capital and materials, or banks use a clearing house and settlement bank to execute the transactions. The latter is particularly important with respect to giro payments, card payments and ATM transactions. The banks are charged for their share of the costs in these systems, and these costs are included in our measure of materials input.

We estimate a system consisting of a cost function in four (primary) variable inputs, i.e. labour, physical capital, materials and funding, and the corresponding cost-share equations. We treat loans and deposits as banks' output within a single-output approach, i.e. loans and deposits are aggregated. Rather than representing technical change by a deterministic trend variable, we include the development in electronic payments relative to total non-cash payments in volume terms, i.e. number of transactions, to represent the implementation of this new technology in the payment systems. We specify a rather general model in this variable, and in addition to testing for an effect on banks' scale elasticity, we also test if new payment systems have affected input demand asymmetric, i.e. non-neutral. These are the two issues that we focus on.

We have access to annual bank-specific data for most of the variables in our system over the period 1987–98. With respect to electronic payments, we only have the number of transactions at the industry level, however, while we ideally should have had the number of transactions at the bank level. On the other hand, we expect the share of electronic payments in total non-cash payments measured in number of transactions to be rather common across banks. Because banks participate in a jointly owned system for clearing, and the necessary infrastructure is in general available to all banks, banks can rather easily offer their customers electronic payment services. For example, the first Norwegian bank to offer e-banking was a small rural savings bank.

Section 2 presents the econometric model, and the empirical results are presented in Section 3. The main conclusions are summarized in Section 4.

2 The cost-share equation system

Banks' technology is represented by a translog cost function with variable returns to scale as put forward by Christensen *et al*. (1971, 1973). With this functional form, the impact of technical progress, as represented by the introduction of new electronic payment systems, can be specified in a general way. The translog cost function is flexible and can be interpreted as a quadratic approximation to a general continuous twice-differentiable cost function in logs that satisfies linear homogeneity in prices.

While Hunter *et al*. (1990) and Lawrence (1989) find that the standard translog specification provides an adequate fit of bank cost data, Shaffer (1998) shows that the translog cost function disfavours large banks, since economies of scale are exhausted already for relatively small banks. In principle, however, the general translog cost function captures situations with ever increasing scale elasticity with respect to bank size. The conclusion with respect to this will depend on the sign and size of estimated coefficients in combination with data. One may therefore argue that this is an empirical question, and if increasing or non-decreasing scale elasticity with respect to bank size is an important feature of the data, we should be able to identify this by using the translog functional form.

Our most general translog cost function, with labour (L), physical capital (K), materials including energy (M) and funding (F) as inputs, is given in equation (1). Subscript f denotes bank. The cost function includes fixed effects, which is assumed to capture permanent variation in cost efficiency across banks. All slope coefficients are treated as constant across banks. The symbol Σ_i implies the sum over all variable inputs.

$$\begin{aligned} \ln C_{\mathrm{f}} = {} & \alpha_{0\mathrm{f}} + \Sigma_i\, \alpha_i \ln P_{i\mathrm{f}} + 1/2\, \Sigma_i \Sigma_j\, \beta_{ij} \ln P_{i\mathrm{f}} \ln P_{j\mathrm{f}} + \gamma_X \ln X_{\mathrm{f}} + 1/2\, \gamma_{XX} (\ln X_{\mathrm{f}})^2 \\ & + \Sigma_i\, \gamma_{iX} \ln P_{i\mathrm{f}} \ln X_{\mathrm{f}} + \gamma_E \ln EP + 1/2\, \gamma_{EE} (\ln EP)^2 + \Sigma_i\, \gamma_{iE} \ln P_{i\mathrm{f}} \ln EP \\ & + \gamma_{XE} \ln X_{\mathrm{f}} \ln EP + u_{C\mathrm{f}} \qquad (1) \\ i,j = {} & L, K, M, F \\ C_{\mathrm{f}} = {} & \Sigma_i\, P_{i\mathrm{f}} \cdot V_{i\mathrm{f}} \qquad i = L, K, M, F \qquad (2) \end{aligned}$$

where C_{f} is total variable costs of bank f; $P_{i\mathrm{f}}$ is the bank-specific price of input i; X_{f} is the output, i.e. the sum of loans and deposits, in bank f; EP is the share of electronic payments in total non-cash payments in the banking industry; $u_{C\mathrm{f}}$ is an added disturbance; $V_{i\mathrm{f}}$ is the quantity of input i used by bank f.

Application of Shephard's lemma gives the cost-share ($S_{i\mathrm{f}}$) equations in (3).

$$\begin{aligned} S_{i\mathrm{f}} = \partial \ln C_{\mathrm{f}} / \partial \ln P_{i\mathrm{f}} &= (P_{i\mathrm{f}} \cdot V_{i\mathrm{f}}) / C_{\mathrm{f}} \\ &= \alpha_i + \Sigma_j \beta_{ij} \ln P_{j\mathrm{f}} + \gamma_{iX} \ln X_{\mathrm{f}} + \gamma_{iE} \ln EP + u_{i\mathrm{f}} \qquad (3) \\ i, j = L, K, M, F & \end{aligned}$$

where $u_{i\mathrm{f}}$, $i = L, K, M, F$, are disturbances. Because the cost shares always sum to unity, that is $\Sigma_i S_{i\mathrm{f}} = 1$, any cost-share equation can be expressed in terms of the other equations by using the adding up restrictions (listed below). This also implies a singular error-covariance matrix, but estimation may proceed with the arbitrary deletion of one cost-share equation.

Theory requires the cost function to be homogeneous of degree one in input prices and cross-price effects to be symmetric. These theoretical restrictions, in addition to the adding up conditions, are imposed on the general model that we estimate:

Adding up conditions:	$\Sigma_i\alpha_i = 1$; $\Sigma_i\beta_{ij} = 0$ for all j;
	$\Sigma_i\gamma_{ik} = 0$ for $k = X, E$
Price homogeneity and symmetry:	$\Sigma_j\beta_{ij} = 0$ for all i;
	$\beta_{ij} = \beta_{ji}$ for all i and j where $i \neq j$.

We make the following assumptions about the disturbances in the reduced system that we estimate (we exclude the cost-share equation for materials):

$$[u_{C\mathrm{f}}, u_{L\mathrm{f}}, u_{K\mathrm{f}}, u_{F\mathrm{f}}]^{/} \sim \text{IIN}\ [0, \Omega], \text{ for all } f$$

where IIN signifies independently, identically, normally distributed. Beyond symmetry, there are no restrictions imposed on the covariance matrix, Ω. The genuine errors are assumed to be homoskedastic across industries and not autocorrelated within industries. We assume that all prices and the share of electronic payments in total non-cash payments are weakly exogenous. We check the validity of treating output as an exogenous variable by using an instrumental variable approach.

The own- and cross-price elasticities of factor demand are given in (4) and (5). These elasticities are defined as the Slutsky analogues, i.e. as output-constrained price elasticities of input quantities. The price elasticities depend on data as well as estimated coefficients, and will in general vary both over time and across banks.

$$\varepsilon_{ii\mathrm{f}} = \beta_{ii}/S_{i\mathrm{f}} + S_{i\mathrm{f}} - 1 \qquad \text{for all } i. \tag{4}$$

$$\varepsilon_{ij\mathrm{f}} = \beta_{ij}/S_{i\mathrm{f}} + S_{j\mathrm{f}} \qquad \text{for } i \neq j \tag{5}$$

Grant (1993) shows that the elasticities of substitution in the translog function case may be evaluated at any expansion point, including points of sample means, as long as the theory restrictions of price homogeneity and Slutsky symmetry hold. The cross-price elasticities are in general not symmetric. The translog cost function has been criticized because the area where the regularity conditions are met can be narrow. Particularly, own-price elasticities have attained focus, because positive values – even within sample – have been revealed. We check the within sample properties, however.

The elasticity of scale, which equals the inverse elasticity of costs with respect to output, and the elasticity of costs with respect to electronic payments, are defined in (6) and (7) respectively. As with the price elasticities, these elasticities depend on both data and estimated coefficients.

$$\varepsilon_X = (\partial \ln C_\mathrm{f}/\partial \ln X_\mathrm{f})^{-1} = (\gamma_X + \gamma_{XX} \ln X_\mathrm{f} + \Sigma_i \gamma_{iX} \ln P_{i\mathrm{f}} + \gamma_{XE} \ln EP)^{-1} \tag{6}$$

$$\varepsilon_E = \partial \ln C_\mathrm{f}/\partial \ln EP = \gamma_E + \gamma_{EE} \ln EP + \Sigma_i\, \gamma_{iE} \ln P_{i\mathrm{f}} + \gamma_{XE} \ln X_\mathrm{f} \tag{7}$$

As already explained, the translog cost function is criticized because it disfavours large banks with respect to scale elasticities. However, the formula for the scale elasticity above shows that this is only true if $\gamma_{XX} > 0$. If $\gamma_{XX} < 0$ or $\gamma_{XX} = 0$, the scale elasticity will increase with, or be independent of, bank size respectively. We expect ε_E to be negative, i.e. that the increase in the share of electronic payments in total non-cash payments has decreased average costs.

There are a number of hypotheses concerning the properties of the production function that can be tested on the general model. We are, however, particularly interested in (i) the effect on the scale elasticity from the increased use of electronic payments, and (ii) if new technology, as represented by new electronic payment systems, has affected input demand asymmetric, i.e. non-neutral. With respect to the first issue, if the coefficient γ_{XE} is zero, which means that no interaction term between output and electronic payments enters the cost function, we conclude that the increase in electronic payments has not affected the scale elasticity in the banking industry. It would also imply that the effect of electronic payments on variable costs is independent of the level of output. If, on the other hand, γ_{XE} is found to be negative (positive), this implies an increase (decrease) in the elasticity of scale as the share of electronic payments increases.

The issue of neutrality can be analysed by testing restrictions on the coefficients γ_{iE}, $i = L, K, M, F$. If, for all i, $\gamma_{iE} = 0$, we conclude that the increased use of electronic payments has affected input demand neutrally, i.e. input ratios are not affected by this change in technology, although the input volumes may of course change. If $\gamma_{iE} \neq 0$, the conclusion is more complicated. Let's assume, however, that the elasticity of costs with respect to electronic payments, ε_E, has the expected negative sign. This implies that an increase in the share of electronic payments has a negative effect on average variable costs and hence an overall negative effect on the volume of variable inputs. If $\gamma_{iE} < 0$, we can conclude that the volume of input i decreases as the share of electronic payments increases, when all prices and the output level remain constant. If, on the other hand, $\gamma_{iE} > 0$, then we have two opposite effects on the demand for input i from an increase in electronic payments. The net effect depends on whether this latter positive effect on the cost share of input i dominates the negative effect on variable costs from ε_E.

Within the chosen approach, we can calculate the effect from increased use of electronic payments on input ratios by using equation (8). ΔIR_{ij} denotes the change in the input ratio between input i and j.

$$\Delta IR_{ij} = \partial(S_{if}/S_{jf})/\partial \ln EP = (\gamma_{iE}\mathrm{S}_{jf} - \gamma_{iE}\mathrm{S}_{if})/(\mathrm{S}_{jf})^2 \qquad (8)$$

If ΔIR_j is positive (negative) we conclude that the input ratio between input i and j increases (decreases) as the share of electronic payments increases. Since electronic payment services are assumed to replace more expensive labour-intensive payment services, we are particularly interested in checking the impact of electronic payments on the input ratios between labour and physical capital and between labour and materials. Electronic payments are – to some degree – contingent on investments in new machines. Furthermore, usage of the banks' jointly owned system for clearing and liquidity information, which clearly is important with respect to electronic payments, is included in materials input.

If all cost shares are independent of the output level, i.e. $\gamma_{iX} = 0$, we conclude that the production technology is homothetic, and factor ratios remain constant when the level of output changes. Homotheticity in addition to the absence of price effects in the cost-share equations, i.e. $\beta_{ij} = 0$ for all i, j, imply a Cobb–Douglas production technology. As in Jorgenson (1986), we define a positive (negative) effect of output growth on a cost share as a positive (negative) scale bias.

3 Empirical results

We now present the results from estimating the cost function in (1) and the cost-share equations in (3). We exclude the cost-share equation of materials for reasons explained earlier. We use bank-level panel data from Norwegian banks combined with industry-level information from national accounts. A more detailed presentation of the data and empirical variables is given in Appendix 15A. Our panel includes 1873 annual observations of 211 banks over the period 1978–98. Our panel is unbalanced, but 141 banks are observed in the maximum 12 years, which is 90 per cent of the observations. Three banks are observed in one year only. We use the three-stage least squares procedure in STATA Release 7.0 (StataCorp., 2001) to estimate the simultaneous system with cross-equation restrictions.

As in general in cost function analyses, we may face an endogeneity problem with respect to the explanatory variables. Wages vary across banks, which may suggest that banks are not price takers in the labour market. However, this variation may reflect differences in seniority and level of education, and one can argue that due to a centralized wage formation system and relatively high degree of unionization in Norway, it is plausible to assume that wages are weakly exogenous. Because we only have access to the industry-level price on materials input (including energy), any possible endogeneity problem with respect to this price is reduced. The price of funding is a weighted average of interest rates on deposit accounts and a money market interest rate. The latter interest rate is clearly exogenous to the banks. We assume that the money market is the 'true' marginal source of funding for the banks. Although it can be argued that the implementation of new technology in banks in general should be treated as an endogenous process, two arguments can be put forward that defend the weak exogeneity assumption on the share of electronic payments in total non-cash payments. First, we have only access to the industry-level share, which is exogenous to each bank, and second, the change in this share is determined not only by the banks but also largely by the bank customers.

To take into account a possible endogeneity problem with respect to output, we use a four-stage least squares (4SLS) approach. In this case, we first regress output on all the exogenous variables in our general system in addition to the volumes of the variable inputs. The predicted values of output from this regression are used as instruments when estimating the system of interest by the three-stage least squares (3SLS) procedure in STATA.

Table 15.1 gives the results from estimating our general model by 3SLS and 4SLS.

Table 15.1 The estimated cost function and cost-share equations [a]

	Three-stage least squares		*Four-stage least squares* [b]	
Coefficient	*Estimate*	*Std. error*	*Estimate*	*Std. error*
α_L	.240	.019	.339	.020
α_K	.004	.009	–.002	.010
α_F	.676	.026	.692	.027
β_{LL}	.010	.003	–.009	.003
β_{LK}	–.005	.001	–.005	.001
β_L	–.012	.003	–.008	.003
β_{KK}	–.003	.001	–.003	.001
β_{KF}	.032	.001	.036	.001
β_{FF}	–.115	.005	–.146	.005
γ_X	–.376	.156	–.171	.143
γ_{XX}	.074	.010	.077	.010
γ_{LX}	–.005	.001	–.005	.001
γ_{KX}	.001	.0003	.001	.0003
γ_{FX}	.009	.002	.010	.002
γ_E	–.068	.085	–.342	.078
γ_{EE}	–.662	.028	–.825	.025
γ_{LE}	.057	.003	.063	.004
γ_{KE}	.022	.002	.025	.002
γ_{FE}	–.219	.007	–.244	.006
γ_{XE}	–.021	.005	–.018	.005
	Three-stage least squares		*Four-stage least squares*	
	RMSE	R^2	*RMSE*	R^2
Cost function	.1406	.9914	.1212	.9936
Labour [c]	.0426	.4463	.0434	.4252
Physical capital [c]	.0214	–.2547	.0220	–.3236
Founding [c]	.0986	.1407	.1052	.0230

Notes:

a The cost-share equation of materials (incl. energy) is excluded. Fixed effects are included in the regressions but not reported in the table.

b In a first step, output is regressed on all the explanatory variables in the cost function in addition to the level of all the inputs. For this single equation: RMSE = 0.0910 and R^2 = 0.9967. Predicted values on output are used as regressors in the three-stage lest squares procedure.

c The cost-share equation of the input stated.

A comparison of the 3SLS and 4SLS regressions in Table 15.1 shows that the estimated coefficients are, in general, relatively robust whether we treat output as an endogenous variable or not, but there are some differences. The two alternative regressions are robust with respect to the two hypotheses of particular interest, however.

First, the coefficient γ_{XE} is found to be significantly negative, which, as already explained, implies that the elasticity of scale increases as the share of electronic payments increases. Second, neutrality is rejected, i.e. the coefficients $\gamma_{iE} \neq 0$, $i = L, K, M, F$. We conclude that the move towards electronic payment systems has caused the cost-shares of both labour, physical capital and materials to increase, while the cost-share of funding has decreased. According to Table 15.3, the elasticity of costs with respect to electronic payments, ε_E, is negative, and hence we can conclude that the isolated effect on the demand for funding from the increase in electronic payments is negative. We also find that $\Delta IR_{LK} < 0$ and $\Delta IR_{LM} < 0$, which imply that the input ratio between labour and both physical capital and materials (including energy) decreases as the share of electronic payments increases. This is consistent with our *a priori* beliefs; new electronic payment systems have particularly substituted out paper-based and labour-intensive methods, cf. Vesela (2000). The results do not support homotheticity, and we find a negative scale bias for both labour and materials (including energy). The scale bias for physical capital is positive.

In Table 15.2 we present the own-price and cross-price elasticities predicted by the model in Table 15.1. The elasticities are calculated at the overall empirical sample means for the cost shares.

Table 15.2 Own- and cross-price elasticities calculated at the overall empirical mean of the cost shares

Own-price elasticities	*Estimate*		*Cross-price elasticities*	*Estimate*	
	3SLS	*4SLS*		*3SLS*	*4SLS*
ε_{LL}	–0.778	–0.912	ε_{LK}	0.005	0.008
ε_{KK}	–1.018	–1.025	ε_{LM}	0.165	0.269
ε_{MM}	–1.614	–1.939	ε_{LF}	0.607	0.635
ε_{FF}	–0.477	–0.523	ε_{KL}	0.025	0.032
			ε_{KM}	–0.436	–0.524
			ε_{KF}	1.430	1.517
			ε_{ML}	0.217	0.354
			ε_{MK}	–0.180	–0.216
			ε_{MF}	1.577	1.801
			ε_{FL}	0.133	0.139
			ε_{FK}	0.091	0.096
			ε_{FM}	0.253	0.288

All own-price elasticities evaluated at sample means have the correct sign. With the exception of one value of ε_{LL} in the 3SLS case, this is true for all sample points in our data. Hence, the results satisfy in general the 'concavity in prices' condition. On average, the demand for materials (including energy) is elastic, while the demand for labour and particularly funding are inelastic. The own-price elasticity for physical capital is close to minus one. On average, most inputs are substitutes in demand. The exception is physical capital and materials, which on average are complements.

Table 15.3 gives the elasticity of scale, cf. equation (6), and the elasticity of costs with respect to the share of electronic payments in total non-cash payments, cf. equation (7). We find a scale elasticity close to one on average; above one when using 3SLS and just below one when using 4SLS. The results also show that an increase in the share of electronic payments in non-cash payments decreases costs by about 0.5 per cent, i.e. over time, the elasticity of scale in banks' production of loans and deposits has increased due to the implementation of new technology in the payment systems.

Table 15.3 The elasticity of scale and elasticity of costs with respect to electronic payments

	3SLS	*4SLS*
ε_X	1.196	0.918
ε_E	–0.414	–0.557

Note: Calculated at the overall empirical means of the cost-shares.

4 Conclusions

This paper analyses the importance of new technologies, i.e. new electronic payment systems, for the development in banks' scale properties and input demand in the production of loans and deposits. A four-factor translog cost function and the corresponding cost-share equations are estimated simultaneously using unbalanced bank-level panel data for the period 1987–98. The inputs are labour, physical capital, materials including electricity and funding. We treat the sum of loans and deposits as an aggregate output. Fixed effects are introduced to capture systematic bank heterogeneity in costs. Output is instrumented using predicted values from a single equation regression.

To represent the implementation of new technology in the payment systems in the Norwegian banking industry, we include the development in the share of electronic payments in total non-cash payments measured in number of transactions. A rather general model is specified in this variable.

The results show that the elasticity of scale has increased as the share of electronic payments has increased. Average variable costs have decreased. The move towards electronic payment systems has affected input demand asymmetric, i.e. non-neutral, causing the cost-shares of both labour, physical capital and materials to increase, while the cost-share of funding decreases. Calculations on the estimated models show that the input ratios between labour and both physical capital and materials (including energy) decrease as the share of electronic payments increases. This is consistent with our *a priori* beliefs; new electronic payment systems have particularly substituted out paper-based and labour-intensive methods.

Appendix 15A: The data and definition of variables

Primarily we use annual bank-level data which banks are obliged to report to the central bank. This is combined with bank-level information on employment from Statistics Norway. In addition, we apply some price indices at the industry level from Statistics Norway, i.e. we use data for industry 63: Bank and insurance. Our panel includes 1,873 annual observations of 211 banks over the period 1987–98. Our panel is unbalanced, but 141 banks are observed in the maximum 12 years, which is 90 per cent of the observations. Three banks are observed in one year only.

C_{f} Total variable costs in bank f, 1000 NOK
$P_{L\mathrm{f}}$ Price of labour in bank f, NOK per man-hour
$P_{K\mathrm{f}}$ User cost of physical capital in bank f, NOK
$P_{M\mathrm{f}}$ Price of materials incl. energy in bank f, NOK 1997 = 1
$P_{F\mathrm{f}}$ Interest rate on funding in bank f
X_{f} Total output in bank f measured as the sum of deposits and loans, 1000 1997–NOK
EP Share of electronic payments in total non-cash payments in the Norwegian banking industry.

Total variable costs are calculated as the sum of reported labour costs, input of physical capital, materials input and funding costs.

We calculate data on two physical capital stock components, i.e. machinery, fixtures and transport equipment, and buildings. The two components are aggregated to total physical capital stock. Data in value terms are the sum of deflated book values multiplied by a calculated user cost. We include rented capital in our measure. The user cost of physical capital is a weigted average of the user cost of each component. Each user cost is a function of the corresponding industry level investment price from the Norwegian national account, the one period ahead rate of change in this price, the ten years Norwegian government bond rate, and a depreciation rate. The latter is based on information on the service life from Norwegian national accounts and geometric depreciation. The service life is nine years for machinery etc. and sixty years for buildings. The user cost is common across banks.

The price of materials including energy is a weighted average of the industry level price of materials, electricity and fuels from Norwegian national accounts. This price is common across banks. Funding costs equal banks' interest payments on deposits and money market funding. The price of funding is – for each bank – calculated as the weighted average of the bank's interest rate on deposits and a money market rate, using ordinary deposits and loans in the money market as weights.

Loans and deposits in volume terms are calculated by deflating observed data in value terms by the consumer price index (CPI). With respect to business loans, one may argue that the GDP-price index is a more appropriate deflator, but the GDP-price index and CPI follow a common trend. With respect to electronic payments, we only have the number of transactions at the industry level, however, while we ideally should have had the number of transactions at the bank level. Table 15A.1 gives summary statistics for the variables used in the regressions. According to the empirical means, the input share of funding is well above the other variable inputs.

Table 15A.1 Summary statistics for the endogenous and explanatory variables in the analysis

	Mean	*St. Dev.*	*Min.*	*Max.*
Total costs, mill. NOK	466	1,655	869	22,700
Cost share of labour	0.154	0.057	0.005	0.581
Cost share of physical capital	0.038	0.019	0.0002	0.193
Cost share of materials including energy	0.137	0.054	0.009	0.745
Cost share of funding	0.671	0.106	0.002	0.963
Price of labour, NOK per man-hour	209	505	5	19,054
User costs of physical capital	0.164	0.076	0.049	0.336
Price of materials including energy[1]	0.899	0.060	0.798	1.027
Interest rate on funding	8.835	16.555	0.978	561.662
Output, mill. 1997–NOK	7,192	27,200	17	292,000
Share of electronic payments[2]	0.431	0.192	0.156	0.741

Notes:
1 Annual national account data, 1997 = 1. Equal for all banks.
2 The share of electronic payments in total non-cash payments in the banking industry. Equal for all banks.

References

Christensen, L.R., Jorgenson, D.W. and Lau, L. (1971) 'Conjugate duality and the transcendental logarithmic production function', *Econometrica* 39: 255–6.

Christensen, L.R., Jorgenson, D.W. and Lau, L. (1973) 'Transcendental logarithmic production frontiers', *Review of Economics and Statistics* 55: 28–45.

Flatraaker, D.-I. and Robinson, P. (1995) 'Income, costs and pricing in the payment system', *Norwegian Central Bank Economic Bulletin* 66: 321–32.

Grant, J.H. (1993), 'The translog approximate function. Substitution among inputs in manufacturing evaluated at sample means', *Economics Letters* 41: 235–40.

Humphrey, D.B., Kim, M. and Vale, B. (2001) 'Realizing the gains from electronic payments: costs, pricing, and payment choice', *Journal of Money, Credit and Banking* 33: 216–34.

Hunter, W.C., Timme, S.G. and Yang, W.K. (1990) 'An examination of cost subadditivity and multiproduct production in large US commercial banks', *Journal of Money, Credit and Banking* 22: 504–25.

Jorgenson, D.W. (1986) 'Econometric methods for modelling producer behaviour' in Z. Griliches and M.D. Intriligator (eds), *Handbook of Econometrics Volume III,* Chapter 3. Amsterdam: North-Holland.

Lawrence, C. (1989) 'Banking costs, generalized functional forms and the estimation of economies of scale and scope', *Journal of Money, Credit and Banking* 21: 368–79.

Shaffer, S. (1998) 'Functional forms and declining average costs', *Journal of Financial Services Research* 14: 91–115.

StataCorp. (2001) *Stata Statistical Software: Release 7.0,* College Station, TX: Stata Corporation.

Vesela, J. (2000) *Technological Transformation and Retail Banking Competition: Implications and Measurement*, Bank of Finland Studies E:20.

Wells, K. (1996) 'Are checks overused?', *Federal Reserve Bank of Minneapolis Quarterly Journal Review* 20: 2–12.

Acknowledgements

The author wishes to thank Bent Vale, Martin Hellwig, Grete Øwre, Aris Kaloudis, and participants at the 23rd SUERF Colloquium in Brussels, October 2001 for valuable comments and discussions.

The views expressed are those of the author and not necessarily those of the Central Bank of Norway.

16 Market structure, innovation and the development of digital money*

Peter D. Spencer

1 Introduction

Like many others attending this colloquium, I am very interested in the implications of electronic money for financial stability and monetary policy. However such analysis begs the question of whether the development and diffusion of electronic money is actually likely. This paper looks at the prospects for the successful diffusion of e-money systems. I look at the obstacles that have to be surmounted and discuss ways in which the competition and prudential authorities could encourage or impede this development.

We have of course had electronic money for decades. Electronic Funds Transfer (EFT) has been used for settling large banking and commercial transactions since the middle of the nineteenth century, based, like many commercial procurement operations, on the telegraph and telex (Standage, 1998). Closer to home, online credit and debit card systems are also electronic. Credit cards are extensively used on the Internet as well as in the real world, and the debit payment system is expanding rapidly in all developed countries.

However, these transactions are conducted through *closed* commercial networks. The dedicated nature of these systems means that they have high overhead costs with high intrinsic security. In contrast, the Internet is an *open* system. Internet-based systems like Digicash and PayPal have a lower overhead cost but potentially lower security. Electronic purses like Mondex, Geltcarte and Proton avoid system overheads by operating offline, but again are potentially less secure. Like notes and coin in the real world, electronic money is bearer rather than account-based money, subject to the same security risk for the holder.

In this paper I define electronic money (or e-money) as an open or offline digital prepayment system that can be used for settling transactions on the Internet or in the real world or both. This is the electronic equivalent of cash, which also intermediates transactions in a decentralized (or offline) way and must be acquired in advance. The challenge is to develop offline electronic systems that are relatively secure, convenient and cost-effective over a range of micro to macro transactions. As will become clear, this is a challenge to regulators as much as

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

digital scientists. Indeed, I will argue that most of the technical problems have already been surmounted.

It is possible that despite the large overheads, the falling cost of dedicated online electronic services could soon make them as cost-effective as offline e-money media. Indeed, special pricing and security arrangements already allow these systems to be used for high-volume/low-value purchases like metro tickets. This would frustrate the development of offline systems, removing their cost advantage in these areas. However, in this case, online systems would secure many of the resource savings that would otherwise be expected from the offline systems.

So far, the offline electronic systems have utterly failed to penetrate the payments market. This paper looks at the reasons for this failure and asks whether the new systems currently being deployed are likely to fare any better. I argue that the obstacles to the success of these systems are due to market structures rather than technological or psychological barriers. These obstacles stem from the essential characteristics of money as (a) a network and (b) a convenience good. In combination, these features of the payments market raise important public policy issues, meaning that we have to look carefully at the role of the competition and prudential regulators. Ultimately there may be a role for the state in the provision of electronic money, just as there is in the provision of its ancient counterpart.

This paper is set out in the following way. The next two sections look at the network and convenience features of money. Section 4 looks at some of its other attributes and identifies reasons why the first-generation systems (like Digicash and Mondex) performed poorly. It argues that the second-generation systems have remedied most of the technical problems. Subsequent sections suggest that the anti-competitive features of the payments market are nevertheless likely to impede the development of e-money and other new payment systems. The role of the competition authorities and prudential regulator is discussed at this point. The final section offers a conclusion.

2 Money as a network good

Money is a prime example of a '*network good*'. These goods are monetary, language and other communications devices that depend for their effectiveness upon the number of other people using them. As the textbooks say, money has to be 'generally acceptable in settlement of transactions'. The more widely acceptable the better.

Because network goods enjoy a positive consumption externality, they are likely to be under-provided by the market. In this sense network goods are akin to public goods like broadcasting, where consumption by one person does not reduce the amount of the goods available for others. Indeed, network goods are 'super-public goods' because consumption by one person increases the usefulness of the goods to other users. The argument for public subsidy (and perhaps provision) holds *a fortiori*. Malinvaud (1980) provides some basic theory.

This consumption externality leads to a chicken and egg problem: people are reluctant to buy network goods unless their associates have them and they are affordable. This means that promoters must invest huge sums of money in subsidizing appliances and other costs of joining the network before they reach commercial viability. This was the case, for example, with BankameriCard: the first credit card. The Bank of America sent out millions of unsolicited cards in an attempt to reach critical mass, knowing that it would be hit by huge fraud and other costs as a consequence. Similarly, in the UK, it took Barclaycard a decade to turn in its first profit. Katz and Shapiro (1985) provided the first microeconomic model of a network market, showing a market could be established under rational expectations, when consumers and producers look through the short-run inertial costs to the long-run equilibrium.

To set against this, the promoter is likely to enjoy a first mover advantage or incumbency effect once the system is established. This is most likely if it is expensive or inconvenient to switch provider. A notorious example in the technology industries is offered by Microsoft. Ausubel (1991) shows that the credit card market also generates super-normal profits. Having made a large investment in plastic, the credit card companies and the banks have been reluctant to develop a rival payments system. Indeed, Visa and MasterCard are now in court to answer the US Justice Department's allegation that they suppressed competition by abandoning plans for new technologies like smart cards and Internet payment systems. Ginguly and Milne (2001) note that the clearing house at the centre of any payments system is usually organized as a mutual organization by its member banks. This clearly raises competition concerns. Indeed, Cruikshank (2000), a former telecoms regulator, has argued that a payment system is a natural monopoly like a public utility. It is hard for newcomers to enter the market, because it involves a great deal of duplication. It is also difficult to get customers to switch providers. The economics of markets with switching costs are analysed by Klemperer (1995).

It is no coincidence that the intermediary structure of the payments market closely resembles that of telecoms. Indeed, securities dealers and Internet Service Providers (ISPs) have a similar intermediary function. These markets allow one agent or end user (say agent A) to transact or converse with another (B). However, unless A and B have the same intermediary (i.e. bank/phone company/dealer/ISP) they have a five-party structure. In each case agent A transacts through his intermediary (say IA) which settles through a clearing house, network or exchange with B's intermediary (IB) and hence indirectly with B.

The multilateral geometry of this market structure makes it very rich. Take, for example, the payment arrangements: who pays? In the case of debit card systems it is the receiver. But in the case of telecoms it is the caller. In the case of securities and ISPs, both of the ultimate parties pay.

Credit card systems are slightly different, having a rectangular or four-party structure. In this case agent A (the purchaser) settles with the card-issuing bank or 'provider' and agent B (the merchant) with the 'acquirer'. There is no network for different cards, so A and B must be on the same system. Bearer or offline e-money systems are slightly different again, having a triangular or three-party structure. This is because the same firm deals with both end users. Charge cards like American Express operate in the same way. However, questions like the one of 'who pays?' also arise in these systems.

3 Money as a convenience good

These markets have another common feature. The product is a '*convenience good*': wanted not for its own sake but as a way to access other goods and services. Money is the archetypal convenience good. This has several implications. First, it means that money acts like a joint good with the item being purchased. This immediately tells us that the price elasticity of demand is low. If the cost of transactions goes up we can try to switch provider, but this may not be possible (monopoly) or practicable (switching costs). We are not going to change our lifestyle, demand for goods and services or even the level of transactions that we make.[1] This point is particularly worrying when set alongside the monopolistic features of the payments market identified in the previous section. A natural monopoly with a low elasticity of demand clearly raises important public policy issues.

Second, because it is a convenience good, money should be convenient and multifunctional. However, the first-generation digital systems relied upon a lot of expensive and cumbersome hardware. Moreover, they were unifunctional. For example, the first Mondex device was designed for settling small transactions in the real world and Digicash's Ecash product was exclusively for Internet use. The proliferation of different e-money systems and standards was also a handicap, as was the cumbersome and expensive hardware. These systems met with stiff resistance from consumers who were happy with their credit cards, knowing that the provider would guarantee the transaction and pay the cost of misuse, at least in the UK.

4 The prospects for e-money

This all helps to explain the failure of the first-generation e-money systems. They were cumbersome, expensive and unifunctional. Business consultants also tend to emphasize security and psychological factors like consumer caution about distance selling and settlement, fraud and confidentiality. However, from the monetary perspective the real problem is that money is a network good and needs a big investment by the promoter to overcome inertia, just as the credit card did. This investment was simply not forthcoming. Indeed the credit card companies and the banks resisted innovation in this area, in order to protect their investment in the incumbent system. The credit and debit card systems are the dominant forms of payment on the Internet as elsewhere and are likely to remain so for some time.

However, there are reasons why the second-generation systems (like the Visacash e-purse and the PayPal digital transfer or cheque system) are more likely to succeed. First, some big potential drivers are evident. The most obvious is the growth of small-scale repeat business over the web: providing digital products like music and entertainment, hire, gambling and game-plays. The AOL/Time-Warner and Seagram/Vivendi mergers underline the commercial potential for such 'content' provision over the Internet.

The convergence of Internet, television and telephone systems means that there is a huge digital market to be exploited. These industries will require a suitable medium for flexible, small-scale payments. If the plastic card incumbents fail to provide a more effective system for such electronic micro-transactions, this will encourage the development of digital devices. Indeed, the content providers may promote their own payment mechanisms. Mobile phone and satellite broadcasting companies already have a lot of expertise in electronic accounting and billing systems, and these can easily be expanded into vending and other areas using current technology. In Finland you can already use vending machines with your mobile phone, paying at the end of the month through your mobile account.

The irony is that the credit card companies have recognized the potential of this market and the threat of competition from mobile telephone companies just as the US Department of Justice has brought its case against Visa and MasterCard. Visa and Nokia are now market-testing chip cards that combine a credit card (Visa) with an e-purse (VisaCash) for small offline payments and an electronic wallet (containing personal details that facilitate the purchase of goods over the Internet). MasterCard is collaborating with the Finnish telecoms group Sonera on similar mobile systems.

These developments suggest that e-purse systems that require a smart-chip may be able to reside on a person's credit or identity card, effectively piggybacking off current systems. This would reduce cost and increase convenience. They would be useful for offline use as well as micro-transactions. The British government is considering the introduction of chip-based identity cards for asylum seekers, and social-security systems in many countries are looking at the possibility of issuing chip-cards to their clients. These welfare systems could easily be combined with e-purse facilities.

The adoption of pre-pay mobile phones suggests another way in which debit-based systems might achieve take-off. *Inter alia*, these have proved very attractive to those who do not have access to the credit system. This reduces the danger of social exclusion and allows a large expansion of the network, reinforcing the membership externality. E-money systems may also be attractive because they carry limited liability – the maximum loss is restricted to the amount programmed into the card.

Another characteristic of e-purse and other bearer e-money systems that could help in their promotion is that it is much easier to pay interest on these balances than on conventional bearer money units like notes and coin. This is true of any notational money system. E-banks are increasingly relying on the payment of interest as a marketing tool in the deposit market.

These second-generation systems are generally multifunctional, combining systems that can be used in both the real and electronic worlds. Security is enhanced by encryption. The hardware is highly portable, designed for use in future mobile-phone-based commerce. Its price is falling rapidly, to the point at which the credit card providers are offering free smart-card readers to their customers, ideal for use on the Internet. In my view most of the technological problems that beset the first-generation systems have been overcome. But the problem of market structure remains formidable.

5 Market contestability and customer switching costs

Markets should only be regulated when they fail. This can happen if there are agents with market power or externalities; or if there is asymmetric information. Market power is a problem in this context because payment systems are like natural monopolies: duplication is inefficient. The banks that operate these systems are vertically integrated, enjoying classic economies of scale and scope in a wide range of money, credit and other banking markets. The network externality means that we also have to consider the second type of imperfection. In principle, asymmetric information should be less relevant in the settlement of transactions than in long-term financial relationships.

Competition issues have come to the fore in this area recently, largely as a result of bank mergers. Naturally these competition inquiries have focused upon the shares of different banking markets that the merging banks would enjoy. They have also considered the constestability of these markets – the ease with which outside organizations can move in or out of them without incurring irreversible entry or exit costs. On the consumer side, the main question concerns switching costs – the ease with which depositors can switch between banks. Finally, some inquiries, notably the UK's Cruikshank Committee (Cruikshank, 2000), have looked at the operation of the payment system and the clearing house. Many European regulators are conducting Cruikshank-style inquiries at the moment.

Entry, exit and switching costs are naturally very high in the case of a payments system. This means that an external regulator is arguably necessary to simulate a competitive environment. In the absence of such regulation, these systems exhibit classic signs of oligopolistic inefficiency: low rates of innovation and inefficient pricing structures.

As Ginguly and Milne (2001) note, retail payment systems are characterized by a slow adoption of new technology. This paper offers several reasons for this, including the scale of investments in computer systems and the expense of retraining staff. This is the essential basis of the case that the US Department of Justice is bringing against Visa and MasterCard.

It is remarkable that although similar considerations should apply in the case of mobile telephony, they have not held back investment and innovation in that area. The introduction of the new second- and third-generation systems must surely devalue the investment made in the existing systems, even if the rapid expansion of the industry means that some additional capacity is needed. However, incumbents and new entrants alike rushed to bid for the new UMTS licences. These companies clearly believed that the fashionable nature of new mobile devices would ensure a successful take-up.

Entry, exit and switching costs are also significant in the mobile telephone industry, frustrating contestability. Nevertheless, the spectacular success of new entrants against the incumbents at the first-generation stage suggests that an open market can spur on technical adoption, even if such costs are significant. This suggests that in the case of digital payment systems, it will be important for the regulator to ensure open entry to companies with digital security and settlement experience migrating from technology, communication and entertainment (TCE) industries. This should spur innovation and combat the inertial tendencies seen to characterize this sector. Ultimately it may be necessary for a regulator to specify investment levels directly, as in the case of a public utility.

6 The regulation of product and pricing structures for conventional payments media

The second point concerns product and pricing structures. Oligopolistic product ranges normally exhibit bundling or cross-selling features. Their pricing structures usually incorporate cross-subsidies. In combination, these structures can have the effect of stifling competition in products and processes that would otherwise be contestable (like the production and rental of telephone and other appliances that were historically bundled together with the service).

This is another telltale sign of market power abuse. That is because in a fully contestable equilibrium without consumer switching costs, the benchmark cost structure is reflected in the charging structure. If costs and charges are not aligned, new entrants are able to pick off the clientele that finds the benchmark structure more attractive. This leaves the incumbents exposed to a simple form of adverse selection, unravelling the disequilibrium. This is the argument used, for example, by Rothschild and Stiglitz (1970) to analyse a contestable insurance market with asymmetric information.

Payment systems differ from TCE industries because they involve at least three parties (merchant, customer and intermediaries), not just two parties (broadcaster and viewer). Their promoters initially established the credit and debit card systems by subsidizing the merchant equipment and charging merchants a turnover fee. At the same time, they subsidized consumer interest payments. However, this initial structure became fossilized. It is remarkable that the relatively low cost of online credit and debit card transfer is not passed on to the merchant via a reduced fee, as it would be in a contestable system. Indeed, it is now passed to the customer through loyalty points. Having become used to this system, the merchants are arguing that the providers should pay for the cost of the new equipment needed for processing the new chip-cards.

Bank account charges have also adhered to a pattern that is out of line with the cost structure. In equilibrium, we would expect to see banks offering a market rate of interest on their deposits and levying account fees related to the cost of transactions. However in practice we tend to observe zero-interest transactions accounts, with low or negligible transactions charges.[2] Historically, banks have used the 'endowment effect' of zero-interest balances to subsidize money transmission.

In the absence of entry, exit and switching costs, this market would be penetrated by new entrants offering high-interest low-transactions accounts, aligning charges with costs. This would allow them to cherry-pick cash-rich customers with relatively low transactions needs, unravelling the disequilibrium. However, in practice entry, exit or (more likely) consumer switching costs clearly frustrate this outcome. This is why Cruikshank and others have argued that regulators need to pay particular attention to switching costs. He suggested, for example, that bank account numbers, like mobile telephone numbers, should be portable, the property of the user rather than the system operator.

The secular fall in interest rates has given this story another twist. Very few banks are now making a profit on their joint money transmission and deposit banking services. However, they stay in this business because it is the core of their business, the basis for profitable loan, insurance and other cross-selling activities. These profits, rather than the endowment effect, now have to subsidize money transmission.

This situation would also unravel in a contestable market, but there seems little chance of this happening within the current market structure. Nevertheless, there is one way that new entrants could exploit this kind of synergy. If low-cost digital money services were used as a platform for e-banking, insurance and other digital services, this might allow new entrants to make inroads into these other markets, hitting the incumbents on their blind side. Technological and market convergence within the TCE industries and their drive to generate profits from existing billing and other overhead facilities by cross-selling might just make this a reality. Although the payment of interest on e-money balances is now a less effective marketing tool, it may still have some role to play.

These considerations largely concern the regulatory framework that is appropriate for long-run industry equilibrium. However, they are pertinent to the discussion of financial innovation, because there are reasons to think that the current misalignment of cost and charging structures for payment media is retarding the development of new media. The consumer, on convenience or cost considerations, dictates the uptake of new technology products, but it is hard to see how the transactions cost advantage of digital money can assert itself as long as consumers are artificially shielded from the costs of conventional media.

The implication of this line of argument is that the regulator must pay careful attention to the pricing of existing transactions media. Credit card charges would seem to require particular scrutiny. New digital media have found it difficult to displace this product given its current pricing structure, which effectively forces the vendor to give interest-free credit, security and even loyalty points to the purchaser. Another way to help the market to work would be to allow vendors to offer discriminatory prices that reflect this subsidy, but in most countries card providers and competition authorities rule against this.

7 The regulation of networks

I now turn to the second reason for market breakdown: consumption externalities. I have argued that network goods are 'super-public goods' because one person's participation encourages others. This externality could be used to make a case for public subsidy or provision, but this has not been necessary in the case of telecoms and other TCE networks. In these markets, the providers have been quite prepared to look to the long-term equilibrium and invest accordingly. The payments market has been the exception rather than the rule.

The regulatory issues normally emerge once the network is up and running. First, there is the question of technical standards. Many commentators have noted that these are a public good. Sometimes, an open system is developed (Linux software), requiring little regulation. Usually, a closed system wins out, requiring other producers to pay a patent or licence fee at least initially (VHS video, Microsoft's DOS-Windows software).

Collaborative ventures represent a halfway house (e.g. Symbian, the partnership developing the UMTS phone operating system). In this case, the primary role of the regulator is to prevent abuse of market power by incumbents and to ensure reasonable access to new entrants. Ginguly and Milne (2001) suggest that clearing house payment systems are essentially of this type. These are typically not-for-profit organizations run on behalf of their members. Credit card systems are similar. In many countries, these institutions are largely self-regulating. Oversight is provided by the central bank, mainly as a safeguard against systemic risk rather than abuse of market power. But until recently, the competition authorities have steered clear of this area.

8 The regulation of product and price structures for new electronic media

TCE and payments industries all face heavy upfront investment costs. Regular update costs also occur for security and technological reasons. In the case of a public good proper like broadcasting and entertainment, the marginal cost of provision to new subscribers is negligible. In the case of super-public goods like phone systems, negative capacity externality effects are balanced by the positive network externality. This is also a feature of electronic and other security dealing platforms.[3]

These industries combat the inertia effect by reducing the consumer's participation cost. They do this by providing consumer equipment like games consoles, digital television decoders and mobile phones well below cost. They then try to recoup these costs via subscriptions, rentals and high user costs.[4] Like the banks, they also try to extract value by cross-selling.

As I have suggested, security considerations in electronic payment markets seem to favour a debit-based payment system, since this offers limited liability to the holder. It also extends the system to those who are not deemed to be creditworthy. This means putting up cash in advance, typically in a zero-interest balance. This still has a significant opportunity cost. Allowing providers to offer interest on positive balances may help to overcome this problem.

9 The EU electronic money directive

This discussion leads me to suggest that subject to basic considerations of solvency, newcomers should be allowed a high degree of freedom in their product and pricing strategies. Attempts to constrain their marketing strategies by restricting cross-selling and cross-subsidization or imposing interest-rate ceilings could well frustrate development. In this respect I am concerned about the European

Union directive on electronic money. While the US Fed has adopted a laissez-faire approach, the European Commission wants a rigid form of regulation. If this directive is adopted by member states, the risk is that digital money providers will simply migrate to an unregulated US environment.

There is one aspect of the directive that is particularly worrying from a competition angle. It suggests that interest payments on credit balances should not be allowed. Such a rule would clearly prevent promoters exploiting a key advantage of e-cash over physical cash.

Article 3 of the directive requires e-money providers to take and refund deposits in national currency. Many commentators have argued that this is very restrictive and would prevent the development of a private monetary standard. However I am not too concerned about this. The economic theory of networks tells us that unless they are potentially dominant, new entrants to a network have a strong incentive to make their systems inter-operable with the incumbent system (Katz and Shapiro, 1985). In terms of money, this means that new entrants would naturally offer convertibility into government money. Indeed, all of the e-money systems under development at the moment are specified in terms of national currencies.

The need for inter-operability has other important consequences. It gives providers a strong incentive to wait until a common standard is agreed before developing and deploying new systems. If they go it alone, and they fail to set the standard, then they will find themselves at a disadvantage to latecomers who agree a common standard.

This effect operates at the national level too. The monetary authorities in Finland and other Scandinavian countries have taken a lead in promoting innovation in their payments markets. They have, for example, persuaded the banks and the public that bank cheques should be replaced by more efficient giro transfer systems, many of which are electronic. They are seriously concerned about the resource costs of operating the present hybrid system and have considered moving to an all-electronic payments system. However, a major factor that has held them back is the lack of a common digital monetary standard in Europe. They do not want to risk investing in a system that would not be inter-operable with those of their neighbours.

It is interesting that the Monetary Authority of Singapore (MAS) has also looked at these issues, but has come to a different decision. It was surprised at the scale of the operational costs suggested by its investigations, which of course fall upon the private as well as the public sector. Whatever the reason for this, the MAS decided to move to an all-electronic payments system. This plan is now advancing and if it is actually implemented on schedule, notes and coin will be withdrawn and replaced by e-purse and other systems in Singapore by the year 2008.

10 Conclusion

The prospects for the development and deployment of e-money systems now look increasingly favourable. The new systems and business models are more effective than previous ones. The payments market has attracted the attention of technology companies and operators from the TCE industries and some of these have already formed alliances with the credit card companies and the banks.

Nevertheless there remain two formidable economic obstacles to the successful development of e-money. The first is the network effect. This means that the promoter faces a large up-front investment cost. It also makes it risky for providers and even individual countries to 'go it alone'. This is seen as a major obstacle in countries like Finland that are in a good position to adopt an all-electronic system, but has not held back Singapore, perhaps reflecting its different geography.

The second obstacle is the oligopolistic nature of the payments industry. This is generally self-regulating, although the prudential authorities provide oversight in order to safeguard stability. In recent years, many commentators (notably Cruikshank, 2000) have called for the regulation of payment systems by the competition authorities. I have argued that the payments market exhibits some very worrying features, resembling a natural monopoly with a low elasticity of demand. The lack of innovation is particularly noteworthy.

This paper suggests several ways in which the regulator could foster the development of digital money. There is a clear need to ensure open markets; minimize the effect of switching costs; and police the pricing structures of both new and old transaction media. There may be a trade-off here with financial stability, but in my view, this is a risk we have to take in order to encourage innovation.

The Scandinavian countries have shown that the regulator can do much to promote innovation in the payments market. However, in view of the obstacles presented by the industrial structure, this may not be enough to spur the development of digital money. It may take a grand policy initiative to get a new payment system off the ground. Ultimately, money is a public good and if the market fails to do this, the state may have to provide digital money in the same way as it does its ancient relative.

Notes

1 Payment services seem to differ from telecommunications, Internet and share-dealing services in this respect. Although I can find no formal studies of price elasticity, this does appear to be lower in the case of payments services than in these other areas.

2 The interest-bearing current (i.e. transactions) account was invented in Scotland during the free banking regime of the nineteenth century but only began to make inroads into the UK market when interest payments and bank charges were deregulated in 1971. Uptake was encouraged by high interest rates during the 1970s and 1980s. However even in the UK, current (transactions) accounts normally offer a zero or negligible interest rate, while transaction charges are waived provided that the account remains in credit.

3 These are arguably different from payment and TCE systems because the initial investment costs are relatively small. Consumer resistance and security concerns have proved less of a problem than in the case of payment systems.

4 Although, theoretically, this is a disequilibrium phenomenon, consumer resistance to membership costs means that it tends to prevail for a long time.

References

APACS (1999) 'Plastic card fraud'. Online <www.apacs.org.uk>

Ausubel, L.M. (1991) 'The failure of competition in the credit card market', *American Economic Review* 81(1): 50–81.

Basel Committee on Banking Supervision (1996) 'The security of electronic money', report by the Committee on Payment and Settlement Systems, ISBN 92-9131-1197.

Basel Committee on Banking Supervision (1998) 'Risk management for electronic banking and money activities', Working Paper BS/97/122.

Cruikshank, D. (2000) 'Review of banking services in the UK'. Online <www.bankreview.com>

Ginguly, B. and Milne, A. (2001) 'Do we need public policy intervention in UK retail payment systems and if so how?', City University Business School discussion paper, London. Online <www.staff.city.ac.uk/a.k.l.milne/payments.htm>

Goodhart, C.A.E. and Krueger, M. (2001) 'The impact of technology on cash usage', LSE–FMG Discussion Paper #374.

Katz, M.L. and Shapiro, C. (1985) 'Network externalities, competition and compatibility', *American Economic Review* 75(3): 424–41.

Klemperer, P. (1995) 'Competition when consumers have switching costs', *Review of Economic Studies*, 62(10): 515–39.

Malinvaud, E. (1980) *Lectures in Microeconomic Theory*. New York: North Holland Elsevier.

Rothschild, M. and Stiglitz, J.E. (1970) 'Equilibrium in competitive insurance markets', *Quarterly Journal of Economics* 90: 629–49.

Standage, P. (1998) *The Victorian Internet*. London: Weidenfeld & Nicholson.

Part V

Contributions on technology and productivity

17 New economy in Europe – reality or mirage?*

Antje Stobbe

1 Introduction

The term 'new economy' has been in fashion over the past few years and – at the same time – has been widely (mis)used. It has reached almost the same degree and scope of use as the term 'globalization', naming many different phenomena. From a macroeconomic perspective the term 'new economy' suggests a permanent increase in productivity growth and a reduction in structural unemployment with low inflation. At the same time, greater stability of output growth was often cited with respect to the macroeconomic growth 'miracle' in the USA in the 1990s (European Commission, 2000: 19). However, the recent sharp correction in US growth casts some doubts on the hypothesis that higher trend growth goes hand in hand with dampened cyclical movements.

At the same time there is considerable debate over the question whether the increasing diffusion of information and communication technologies (ICT)[1] is a driving force of higher labour productivity.[2] Often, the USA experience during the 1990s, which is characterized by a positive correlation between ICT spending and productivity growth, is taken as proof. However, several empirical studies reveal that most of the productivity effect can be traced to the information technology (IT) producing sector. A rise in total factor productivity (TFP), by contrast, is fairly limited. This puzzle is solved to some extent when the Internet is understood as a general purpose technology (GPT). In that case a large-scale productivity-enhancing effect can be observed only after a time lag when the structural change at company and sectoral level (caused by the increasing use of information technology, i.e. computers, the Internet, wireless technologies) results in higher TFP growth. While the evidence of the 'new economy phenomenon' for the USA is fairly thin so far, it is even more limited for Europe. This article provides evidence of the 'new economy phenomenon' in Europe. Before analysing the macroeconomic development, the article focuses briefly on the findings for the US and the concept of a GPT.

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

2 The findings for the USA

When asking what Europe's position with respect to the new economy phenomenon is, it is helpful to analyse the 'benchmark', namely the USA. In an analysis of productivity growth the relevant measures are labour productivity[3] and total factor productivity (TFP).[4] Labour productivity growth is – *ceteris paribus* – determined by changes in the capital stock (capital deepening) and changes in total factor productivity, representing an increase in overall economic efficiency (technical progress). Thus we can denote the growth rate of labour productivity ($y - l$) as follows,

$$y - l = (k - l) \times wk + tfp$$

where y is the growth rate of output, l is the growth rate of labour input, k is the growth rate of capital stock, tfp is the growth rate of TFP and wk is the weight of capital in production.

Among the studies analysing the development of productivity in the USA, the work of Oliner and Sichel (2000) on the one hand and Gordon (2000a) on the other are the ones most frequently cited. According to Gordon's empirical estimates the increase in (trend) labour productivity in the USA of 0.81 pp in 1995–9 (compared with 1972–95) is mainly attributable to capital accumulation (0.33 pp) and less so to a rise in TFP (0.29 pp). Moreover, in an estimate for the subsectors of the manufacturing sector, the rise in TFP is concentrated in the durable goods sector, suggesting that the production of ICT goods has become more efficient.

By contrast, Oliner and Sichel – and more recently Nordhaus (2001) – show that a larger share of the productivity rise can be attributed to an increase in total factor productivity in the non-farm business sector. Oliner and Sichel estimate that labour productivity increased by 1.05 pp in the period from 1996 to 1999 over the period from 1990 to 1995. This increase is attributable to capital deepening (0.49 pp), but also to a considerable rise in total factor productivity (0.68 pp) which is observed both in the ICT sector and in the rest of the non-farm business sector.

Evidence for a substantial spillover effect from ICT usage to other sectors of the economy, i.e. efficiency-improving effects, are consequently fairly limited. By contrast, most evidence is in favour of the hypothesis that ICT production undergoes substantial productivity-enhancing effects itself.

3 The Internet: a new general-purpose technology?

As shown above, capital deepening has been an important determinant of rising productivity growth even in the USA. The effect of higher ICT investment on total factor productivity remains uncertain, however (ECB, 2001: 39). Nevertheless, this puzzle can be solved to a certain extent when assuming the Internet to be a new general-purpose technology which has the property of resulting in a fundamental restructuring of the economy in general, and production in particular (Speyer, 2000). Under this assumption, the rising usage of the Internet does increase productivity, but only after a time lag.

General-purpose technologies constitute a radical innovation.[5] Whether this is the case for the Internet is still under debate. Recent articles in the *Financial Times* reveal that even among academics there is no common understanding concerning this issue. In a reply to an article by Professor Robert Shiller which had classified the Internet as 'Just another superhighway' (Shiller, 2001) Professor Ian Mackintosh (Mackintosh, 2001) argues that

> the road system was merely an extension and improvement of existing means of physical communication and is, by definition, of limited scope. The Internet, on the other hand, has presented us with a radically new means of electronic communication and is essentially infinite in its economic impact.

While it is certainly debatable whether the impact of the Internet is really 'infinite', the statement reflects the diverging assessment of the two scholars with respect to the Internet being a GPT.

Such an assessment can only be made with certainty in retrospect. However, there are several characteristics of the Internet which let us conclude that it can be classified as a new GPT. First, the Internet certainly has a *wide range of applications*. Being more than simply a (interactive or one-dimensional) communication medium the web displays information, serves as a platform for trade or collaboration and as a distribution channel. Second, it can be *used in many different industries*, be they financial services, retail, automotives, the chemical industry or part of the public sector. Third, the Internet not only itself *undergoes rapid technological change*, but it is also an important driver in the development of *complementary technologies*, be they hardware, software or infrastructure. Fourth, *business processes* undergo substantial reorganization when the Internet is made applicable to them. Virtual marketplaces with a focus on procurement, for example, require companies to computerize every step of their supply chain in order to reap the benefits of online inventory management and ordering. From this example it becomes clear that the Internet is not a technology to be used in itself but is an '*enabling technology*' – one important characteristic of a GPT.

Having thus classified the Internet as a GPT it becomes quite evident that the full productivity-enhancing effects will be felt only with a significant time lag. This is evident from historical examples. But it also makes sense when considering the reorganization efforts which companies have to undergo when employing the Internet, the uncertainty under which they act when assessing the potential benefits of using the net and the replacement costs of existing technologies. Such a delayed effect would also be consistent with the empirical fact that some increase in total factor productivity growth has already taken place in the USA – which is significantly ahead of Europe with respect to ICT spending and Internet diffusion – but not yet on a large scale in Europe.

4 The microeconomic changes in Europe

According to the empirical findings of the studies on the USA, investment in as well as the employment of ICT have the potential of positively affecting labour productivity growth. ICT can influence labour productivity growth through different channels. One is of course capital deepening, i.e. the relative accumulation of ICT capital versus labour. Production becomes more capital intensive and labour productivity rises. A second channel is through spillover effects from the increased diffusion of ICT in the economy. This would result in higher TFP growth. ICT can be employed in various forms: the Internet, mobile services and the use of communication technologies to name a few. The networking of businesses and the leaner allocation of resources in business processes with the help of the Internet should – after a time – positively influence total factor productivity growth. This will be the case if companies are able to reorganize and streamline their production process, e.g. by employing supply-chain management tools, or to optimize customer relationship management (CRM). Moreover, economy-wide network externalities as ICT spreads could contribute to stronger productivity growth. The third channel refers to the rapid technical progress in the production of IT goods themselves which should be reflected in TFP growth. To sum up, both the employment of ICT within, and between, companies and the technical progress in the production of ICT should be driving forces of an increase in the efficiency of the economic process (TFP growth).

While capital deepening is fairly easy to measure by comparing ICT spending across European countries, the increase in TFP is, by definition, not directly measurable. A case could be made, though, that efficiency rises when a certain threshold (tipping point) in the diffusion of IT technology within an economy is crossed. Only then are positive network effects likely to occur: '… information technologies, such as Internet-ready computers, may create network effects that spur the dissemination of information, resulting in disembodied technical change' (Gust and Marquez, 2000: 678). The diffusion of ICT can be measured by the penetration of Internet connections, PCs or mobile subscriptions.[6] However, the available data refer to households and thus do not reflect the employment of ICT technologies on a business level where the efficiency-improving effects should be felt primarily. Because of the lack of cross-country data on the business level we use household data to give a basic idea about the degree to which diffusion of ICT technology varies from country to country within Europe. Moreover, as Internet penetration of households is in some cases complementary to that of businesses – e.g. in the financial services industry where online banking requires households to be online – the following analysis might have more than an illustrative justification.

In an analysis of ICT spending in Europe (capital deepening) it is apparent that there are substantial differences from country to country. Generally, IT spending – less so ICT spending – correlates positively with GDP *per capita* (Figure 17.1). As Figure 17.2 shows, the USA and Switzerland are clearly the leaders when measuring ICT spending *per capita*. Several Nordic countries (Sweden, Denmark, Norway) are next in line, while Italy, Portugal, Spain and Greece bring up the rear.

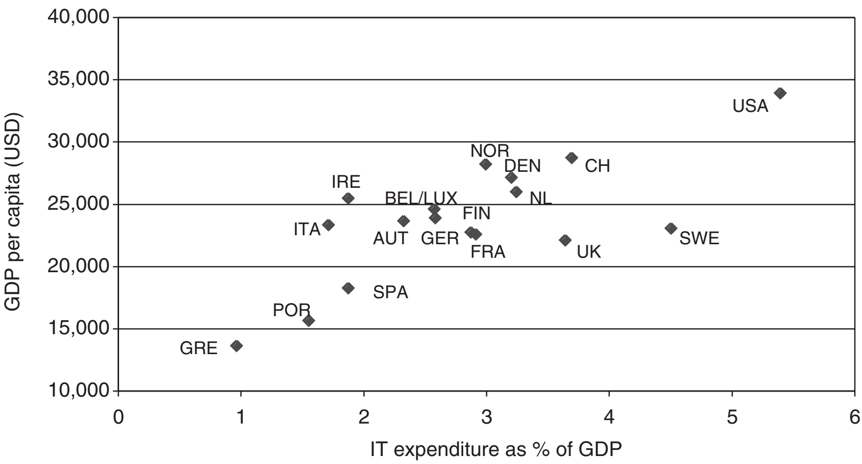


Figure 17.1 IT expenditure and GDP *per capita* (1999).
Source: OECD, EITO, 2001.

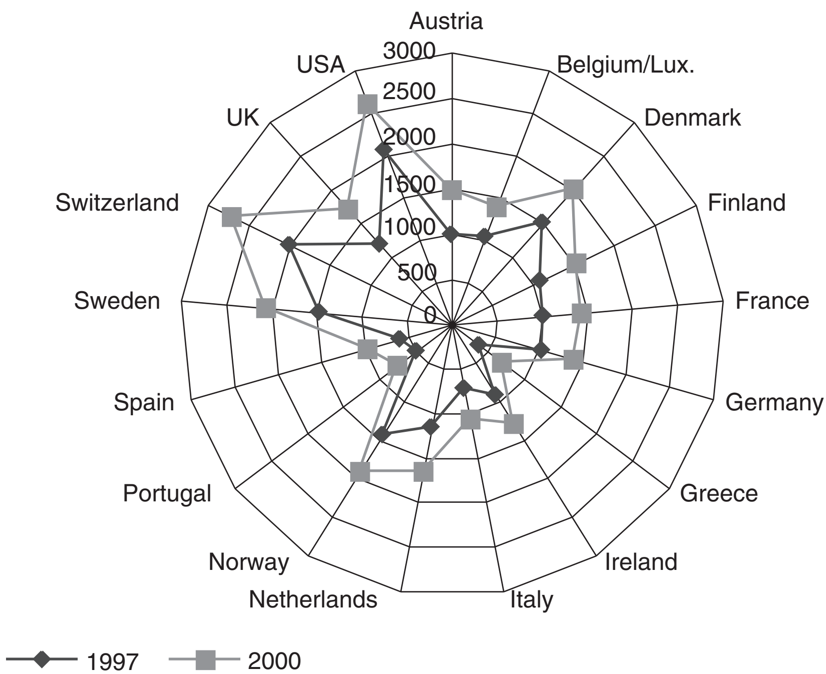


Figure 17.2 ICT spending *per capita* (EUR).
Source: EITO, 2001.

In terms of ICT investment as a percentage of GDP (Figure 17.3), however, the UK also ranks among the most advanced nations. Some of the big European countries like France and Germany are close to the EU average (also Finland).

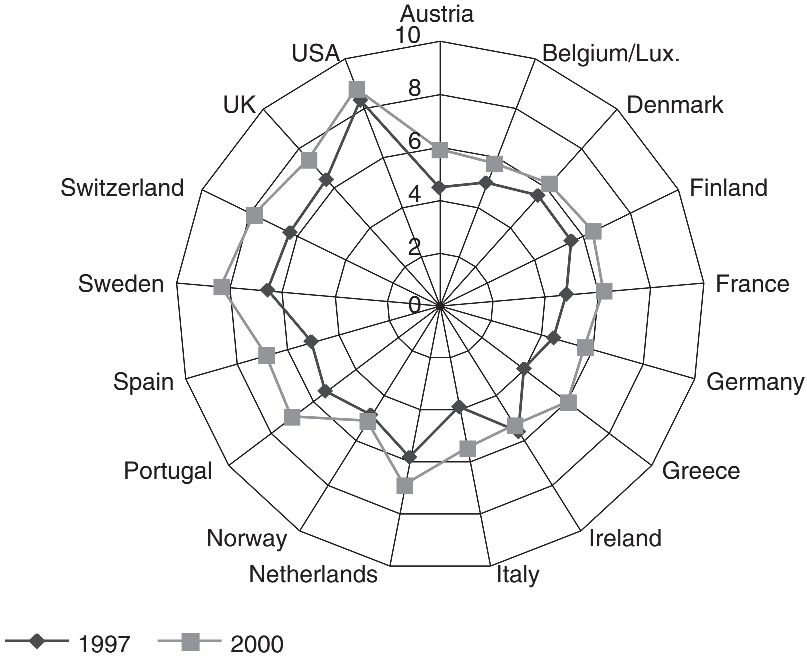


Figure 17.3 ICT spending (% GDP).
Source: EITO, 2001.

It is important to note that a number of those countries that belonged to the laggards in terms of ICT spending in 1997 stepped up their efforts substantially in recent years. This is especially true for Greece, Portugal and Spain with respect to ICT expenditure as a percentage of GDP. However, hand in hand with these countries' comparatively low *per capita* income, ICT spending *per capita* is still significantly trailing that of the leading countries. Also, as calculations by Daveri show (Gros, 2001: 15),[7] even if one assumes the EU average growth rate of ICT spending as a percentage of GDP over the last decade (in Daveri's example, 12 per cent) to persist, it would take almost thirty years for the EU to reach the level of US ICT spending, assuming that the US share of ICT spending in GDP does not increase further. Thus, it becomes clear that the lagging countries have to step up their efforts even more to reach (at least) the US level soon.

The regional gap between the Nordic countries on the one hand and the southern European region on the other also becomes apparent when looking at PC and Internet penetration and the quality of the infrastructure from country to country. With respect to PC penetration (Figure 17.4) the Scandinavian countries are almost as well equipped as the USA, where 60 per cent of the population own a PC. By contrast, in southern Europe roughly 10 per cent of the population own a PC; only in Italy does the ratio come to 20 per cent.

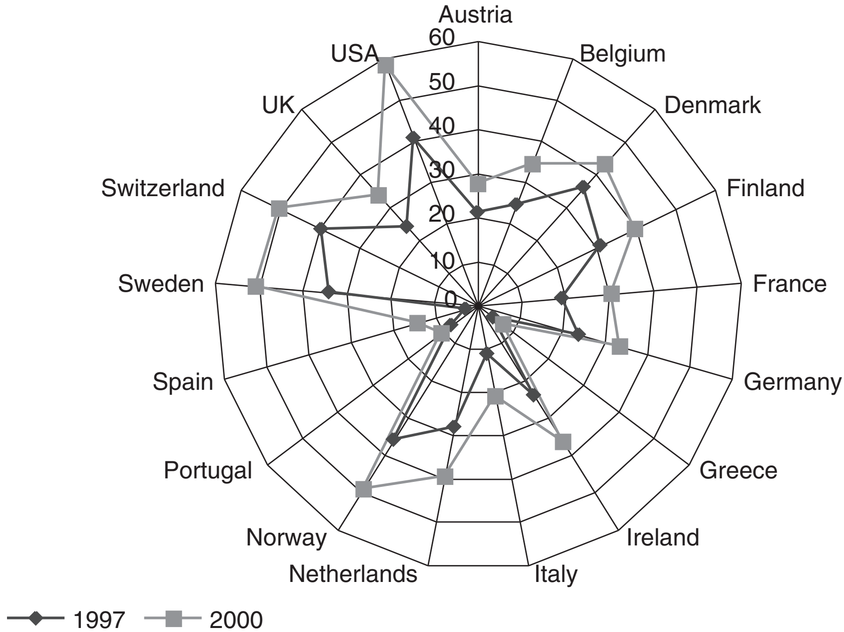


Figure 17.4 PCs (% of population).
Source: ITU.

Although Europeans do not use the Internet as extensively on average as Americans the number of users is growing quite uniformly and very fast across all European regions (Figure 17.5). Sweden and Norway even surpassed the USA's penetration rate in 2000 at around 45 per cent of the population. By contrast, in Italy and Spain only about 10–15 per cent of the population were online that year. Germany, Switzerland as well as the UK were at midfield, with a share of around 30 per cent. The same picture holds true for the number of Internet hosts: northern European countries have on average 10 hosts per 100 inhabitants, versus fewer than 1.5 in the southern European countries and 2.5–3 in Germany or the UK.

Europe as a whole is evidently trailing the USA with respect to the structural change induced by the Internet on the micro-level. Moreover, there is a divide within Europe with respect to the penetration and use of ICT. The Nordic countries appear to be best equipped and have reached or even exceeded the USA level, whereas the southern European countries lag behind. Although it can be noted that some of these countries have increased ICT investment substantially, it will take years until they have reached the penetration level of the top group. The macroeconomic new economy effects in Europe should as a consequence not only diverge substantially from those in the USA, but productivity-enhancing effects should also be different within Europe depending on the level and diffusion of ICT investment. Thus, micro- and macro-analysis of new-economy phenomena on an aggregated European level are misleading. Regional data should be used instead.

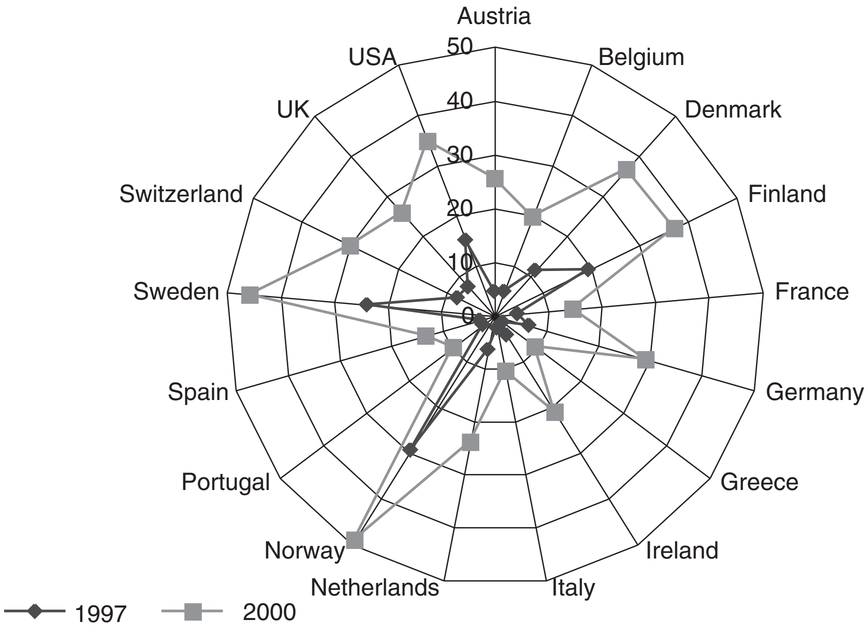


Figure 17.5 Internet users (% of population).
Source: ITU.

5 Macroeconomic new economy effects: now or later?

This section presents some findings of recent studies on productivity growth in Europe. We set investigations into aggregated data – which find very limited influence of ICT spending on labour productivity or total factor productivity – against findings for single countries.

The European Commission (European Commission, 2000) estimates that labour productivity growth has actually slowed to 1.5 per cent per annum in the period 1995–9 compared with 2 per cent in 1991–5 (Table 17.1). The main factor of influence for this slow down is the rise in employment in the European countries in the second half of the 1990s which resulted in a substantial substitution of labour for capital. Another important factor is a reduction in the contribution of non-IT capital (from 0.3–0.4 to 0.1–0.3 pp). The contribution of IT capital to labour productivity (capital deepening), by contrast, increased slightly. At the same time total factor productivity growth in the IT sectors – defined as the computer and semiconductor industries – increased only marginally in the second half of the 1990s (from 0.1 per cent to 0.2 per cent). However, overall TFP growth stagnated as TFP growth in the other sectors of the economy actually decreased, thus cancelling out the small rise in TFP growth in the ICT sector. Consequently, the study by the Commission does not give any hard evidence for the existence of a 'new economy' in Europe on an aggregated level.

Table 17.1 Growth sources in the EU and contribution of ICT sector

	1991–5	*1995–9*	*Change*
Labour productivity growth	2.0	1.5	–0.5
Capital deepening	1.0	0.5	–0.5
ICT capital	0.2–0.3	0.3–0.5	0.1–0.2
capital/labour substitution	0.4	–0.1	–0.5
other capital	0.3–0.4	0.1–0.3	–0.1– –0.2
TFP growth	1.0	1.0	0
ICT sector	0.1	0.2	0.1
other sectors	0.9	0.8	–0.1
Memo ICT share	0.3–0.4	0.5–0.7	0.2–0.3

Source: European Commission, 2000.

However, the Commission's study suffers from several shortcomings. First, it does not include software when analysing the effects of capital deepening – a component which is of significant importance in the USA (Gros, 2001: 12). Second, labour productivity is measured on the basis of individuals, not hours worked.[8] Against the backdrop of the shortening of the work week in numerous European countries over the past years and the rising importance of part-time work calculations on the basis of the number of employees lead to an underestimation of labour productivity and thus TFP growth.

A study by the European Central Bank (ECB, 2001: 37) eliminates the problem described above by calculating euro area productivity growth on the basis of hours worked. Taking into account that average hours worked *per capita* decreased 0.37 per cent in the period 1990–2000, the ECB study still calculates a slight downward bias in productivity growth in Euroland as a whole over the past decade. However, the decline in the labour productivity growth rate is much less pronounced than the Commission's analysis suggests. By contrast, taking only the four big EU countries (France, Germany, Italy and The Netherlands) labour productivity growth (per hour worked) decreased much faster (from 2.4 per cent in 1991–5 to 1.3 per cent in 1996–9) (Table 17.2). The determinants of this development are a decreasing contribution of capital other than ICT to growth and a declining growth rate of total factor productivity compared with the first half of the 1990s. By contrast, an increase in the rate of ICT capital deepening is observable in the second half of the 1990s. Both software and hardware play a major role in capital deepening whereas the contribution of communication equipment is negligible.

Although a more detailed analysis of the sectoral development of ICT spending does not take account of the change in working hours and uses the simpler approach on the basis of employees, it sheds some important light on the contributions to productivity growth by ICT-producing and using sectors. The sector analysis is based on aggregated figures for Germany, France, Italy and Finland (see Table 17.3). It is apparent that – comparable to the development in the USA – the highest growth in labour productivity is observable in the ICT-producing sectors (manufacturing). Labour productivity grew 14.2 per cent in the period 1995–8 compared with 12.9 per cent in 1991–8. Growth in the second half of the 1990s was obviously more

Table 17.2 Decomposition of euro area labour productivity growth

	1991–5	*1996–9*
	Absolute contribution to growth (percentage points)	
ICT capital deepening	0.26	0.39
information equipment	0.10	0.14
software	0.10	0.21
communications equipment	0.06	0.04
Other capital deepening	0.73	0.28
TFP	1.41	0.61
	Annual average percentage growth	
Labour productivity	2.4	1.3

Source: ECB, 2001.

Note: Euro area estimate based on France, Germany, Italy and The Netherlands, which together account for around 77 per cent of euro area nominal gross value added.

Table 17.3 Sectoral developments in the euro area (%)

	Share in nominal value added		*Growth in real value added*		*Growth in employment*		*Growth in labour productivity*	
	1991	*1998*	*'91–8*	*'95–8*	*'91–8*	*'95–8*	*'91–8*	*'95–8*
ICT-producing sectors, manufacturing	0.9	0.7	6.5	11.5	–5.6	–2.3	12.9	14.2
ICT-producing sectors, services	3.6	4.2	5.5	8.1	–0.5	0.1	6.1	7.9
ICT-using sectors, manufacturing	4.5	3.9	0.8	1.6	–3.0	–1.1	3.9	2.7
ICT-using sectors, services	11.3	12.0	2.4	3.2	2.2	2.9	0.2	0.3
Manufacturing	21.0	18.6	0.7	1.5	–2.5	–0.6	3.3	2.1
Business services	47.9	51.8	2.2	2.7	1.0	1.8	1.2	0.9
Total economy	100	100	1.5	1.9	–0.3	0.4	1.8	1.4

Source: ECB, 2001.

Note: Euro area estimate based on Germany, France, Italy and Finland, which together account for around 73 per cent of euro area nominal gross value added.

dynamic, thus also mirroring the US trend. In the USA the figures were 19.2 per cent and 21.3 per cent, respectively (Table 17.4). In contrast to the US data, in the four euro-area countries labour productivity growth was also highly dynamic in the services-producing ICT sectors, picking up substantially in the second half of the 1990s. As for the ICT-using sectors, the results are ambiguous both for the USA and the EU aggregate. The ECB thus concludes that '… the absence of stronger dynamics in the ICT-using sectors than on average in the manufacturing and the business services sectors suggests that, over the period examined, positive spillover effects

Table 17.4 Sectoral developments in the USA (%)

	Share in nominal value added		*Growth in real value added*		*Growth in employment*		*Growth in labour productivity*	
	1991	*1998*	*'91–8*	*'95–8*	*'91–8*	*'95–8*	*'91–8*	*'95–8*
ICT-producing sectors, manufacturing	1.5	1.8	20.9	25.6	1.4	3.5	19.2	21.3
ICT-producing sectors, services	4.0	4.8	6.3	7.8	3.9	5.3	2.3	2.4
ICT-using sectors, manufacturing	3.4	3.0	2.4	2.9	–0.9	0.1	3.3	2.7
ICT-using sectors, services	10.4	13.1	4.7	7.4	3.4	4.5	1.2	2.7
Manufacturing	17.4	16.4	4.5	4.1	0.3	0.6	4.2	3.5
Business services	48.3	52.7	4.8	6.6	2.6	2.9	2.2	3.7
Total economy	100	100	3.5	4.0	1.8	2.0	1.7	2.0

Source: ECB, 2001.

from the use of ICT have only been limited if present at all' (ECB, 2001: 42). However, the sector study by the ECB also suggests that spillover effects for the USA were fairly limited and occurred, if at all, only in the ICT-using services sectors which record an increase in labour productivity growth.

The aggregation of data for different EU countries is certainly justified to get an adequate picture for the region as a whole when assessing 'new economy' effects. This is of major importance for the reaction of monetary policy to a potentially higher 'speed limit' on the growth of the euro-area economy. However the aggregation of figures might actually blur the results because of significantly different ICT spending levels and productivity developments within Europe. The Commission states that 'the aggregate picture masks considerable differences across EU Member States' (EU Commission, 2000: 102).[9] One could make the case that European countries with a significantly higher share of ICT investment as a percentage of GDP and higher penetration rates, e.g. of PCs and Internet, might actually be in a better position to experience new economy effects similar to those in the USA. While only some of the countries in question are euro-area participants – and the results are thus of limited interest to the ECB – the results are nevertheless of importance to get a better understanding of the effects caused by higher ICT spending.

A detailed study of labour productivity in the OECD countries by the Federal Reserve (Gust and Marquez, 2000) shows that next to the USA, Switzerland was the only European country which experienced an acceleration in labour productivity growth in the second half of the 1990s compared with the first (Table 17.5). In most larger European countries (Germany, France) labour productivity growth clearly decelerated, while in some smaller countries with an average share of ICT spending in GDP labour productivity growth rates remained on an elevated level during the 1990s (Finland, Ireland). By contrast, Italy (a clear laggard with

Table 17.5 Productivity estimates 1981–99

Country and productivity estimate	*1981–9*	*1990–8*	*1990–5*	*1996–8*	*1996–9*
France					
Labour productivity	3.41	2.12	2.26	1.86	1.61
Capital deepening	1.10	1.09	1.35	0.57	0.50
TFP	2.26	1.03	0.89	1.31	1.12
Germany					
Labour productivity	n.a.	2.13	2.26	1.96	2.14
Capital deepening	n.a.	1.09	1.22	0.91	1.06
TFP	n.a.	1.03	1.02	1.04	1.07
Italy					
Labour productivity	2.33	2.09	2.72	0.81	0.67
Capital deepening	0.87	1.18	1.36	0.82	0.82
TFP	1.45	0.88	1.32	–0.01	–0.14
United Kingdom					
Labour productivity	3.37	1.72	1.78	1.60	1.47
Capital deepening	0.42	0.53	0.57	0.44	0.54
TFP	2.90	1.20	1.21	1.18	0.95
Belgium					
Labour productivity	2.32	1.90	2.18	1.35	1.05
Capital deepening	0.82	1.06	1.28	0.63	0.60
TFP	1.51	0.83	0.87	0.73	0.46
Denmark					
Labour productivity	2.53	2.67	3.69	0.62	0.86
Capital deepening	n.a.	0.94	1.27	0.27	0.56
TFP	n.a.	1.70	2.37	0.37	0.31
Finland					
Labour productivity	3.85	3.82	3.91	3.66	3.10
Capital deepening	n.a.	n.a.	n.a.	–0.54	–0.53
TFP	n.a.	n.a.	n.a.	4.28	3.70
Ireland					
Labour productivity	5.14	4.01	4.10	3.81	3.96
Capital deepening	n.a.	–0.14	0.15	–0.71	–0.39
TFP	n.a.	4.22	4.01	4.62	4.47
The Netherlands					
Labour productivity	3.40	2.07	2.98	0.23	0.35
Capital deepening	n.a.	0.49	0.90	–0.33	–0.21
TFP	n.a.	1.51	1.99	0.54	0.55
Norway					
Labour productivity	1.44	2.27	3.18	1.80	1.39
Capital deepening	0.92	0.48	0.66	0.12	0.29
TFP	0.50	2.23	2.48	1.73	1.13
Spain					
Labour productivity	3.89	1.96	2.58	0.70	0.34
Capital deepening	n.a.	1.48	2.01	0.40	0.26
TFP	n.a.	0.45	0.52	0.31	0.08

continued

Table 17.5 continued

Country and productivity estimate	*1981–9*	*1990–8*	*1990–5*	*1996–8*	*1996–9*
Sweden					
Labour productivity	1.52	2.06	2.11	1.96	1.73
Capital deepening	0.61	0.81	0.89	0.65	n.a.
TFP	0.92	1.23	1.19	1.32	n.a.
Switzerland					
Labour productivity	n.a.	1.31	0.66	2.38	1.90
Capital deepening	n.a.	1.18	1.21	1.13	1.03
TFP	n.a.	0.10	–0.57	1.20	0.84

Source: Gust and Marquez 2000.

Note: Data for 1996–9 use authors' estimates for labour hours in 1999 for OECD data.

respect to ICT spending), but also The Netherlands and Spain, saw labour productivity growth collapse.[10]

In all European countries a slowdown in the rate of capital deepening was observable which was particularly pronounced in Spain and The Netherlands.[11] However, changes in the rate of multifactor productivity growth were generally of higher importance to changes in labour productivity than capital deepening. Gust and Marquez conclude that 'changes in multifactor productivity growth rates thus appear to have played the pre-eminent role in accounting for divergences in the movements of growth in US and foreign [European] labor productivity' (Gust and Marquez, 2000: 672). However, this observation does not imply that TFP growth strongly correlates with countries' *per capita* ICT spending. While such a case could be made for Switzerland and Sweden (positive correlation) on the one hand and for Italy and Spain (negative correlation) on the other, Norway saw a deceleration in TFP growth even though it is among those countries recording the highest ICT spending *per capita* in a European comparison. The same is true for Denmark. However, one has to acknowledge that Norway, Sweden and Finland report above-average TFP growth rates in Europe.

The results obtained by the European Commission sketch a similar picture. The Commission finds that there are some countries in which TFP growth has risen significantly in the 1990s compared with the 1980s, e.g. in Ireland, Finland and Sweden. With the exception of Ireland, however, TFP growth slowed in the second half of the 1990s, although these countries maintained far above-average TFP growth. For most bigger European countries a slowdown can be observed – with the exception of Belgium and France, where TFP growth improved on very low levels.

Summing up, most EU countries clearly experienced a decline in labour productivity and total factor productivity growth during the second half of the 1990s compared with the first half. New economy effects, i.e. spillover effects of increasing ICT spending, can thus not be observed for the EU as a whole. However, the analysis of country-specific data suggests a slightly different picture. Those countries with high ICT spending *per capita* in most cases also experience TFP growth on a higher level, although it does not accelerate (with the exception of Switzerland). While thus a case can be made for not using aggregate European data when analysing so-called new economy effects as the inhomogenous development of productivity within Europe might lead to misleading results, hard evidence for spillover effects remains fairly weak. This is also the case because detailed data on the contribution of ICT to TFP growth are not available.

Two broad sets of open questions remain. The first deals with the reasons for decreasing labour productivity growth in Europe versus an acceleration in the USA. In this context differences in how prices are measured[12] (methodology) and cyclical arguments[13] are mentioned frequently. The second is actually a subset of the first and deals with the determinants of TFP growth as the analysis of ICT spending has remained insufficient. We concentrate on the latter.

6 What is determining TFP growth in Europe?

Changes in labour quality and spillover effects of a relatively large ICT-producing sector – as observed in the USA – might be factors affecting TFP growth. The first argument essentially captures the adverse effects labour-market policies in Europe might have on TFP growth and explains differences on a country level within Europe as well as differences versus the USA.

The second argument seeks to make a case for TFP growth increasing on account of innovation in the ICT sector itself, thus trying to explain the diverging development of productivity growth versus the USA. Sketching TFP growth and the share of ICT in production for various European countries (Figure 17.6), a positive correlation is observed suggesting that countries with larger ICT sectors indeed experience higher TFP growth. The share of the ICT-producing sector in the USA and Finland was, at 1.8 per cent of nominal value added, more than twice as high as in other European countries under consideration. This might actually explain Finland's extraordinary TFP performance against the backdrop of the country's close to average ICT spending. Regression results calculated by the EU

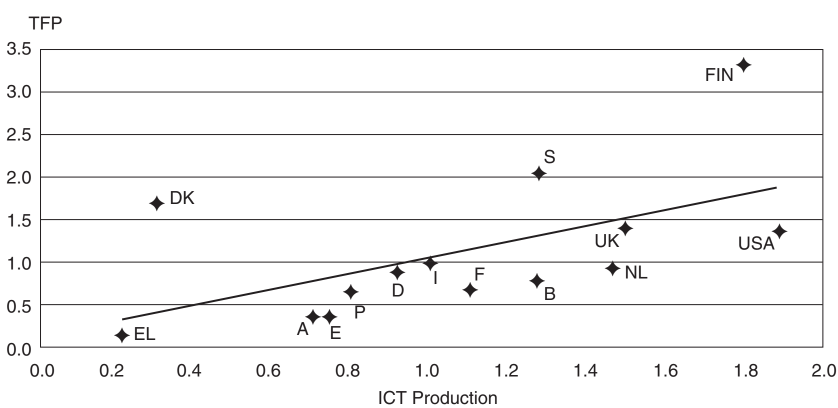


Figure 17.6 TFP growth and ICT production share (1992–8) (%).
Source: European Commission, 2000: 128.

Commission show that the share of ICT production does play a major role in explaining TFP growth for Europe as a whole (European Commission, 2000: 106 and 129). This is not really surprising as it tallies with the results obtained for the USA, that technical progress in ICT production actually plays a major role in the growth of TFP. It also is in line with the results obtained by the ECB.

However, detailed calculations of the contribution of the share of ICT production to TFP growth are considerably flawed by the lack of data on TFP growth for the subsectors of the economy (European Commission, 2000: 219f.). Also, according to the above-cited regression results, the Commission finds that this factor does not fully explain the extraordinary performance of the USA, Sweden and Finland in terms of TFP growth. A dummy employed for these countries in the regression – and thus a not fully explainable factor – actually has the largest coefficient. The Commission concludes that the high significance of the dummy 'is due more to sector specific developments in these countries, as evidenced by the productivity acceleration in these countries' high tech sector'.

The quality of human capital could also be a determinant for the differences in productivity development in Europe versus the USA. An OECD study shows that in the USA labour quality was relatively stable between 1981 and 1998 (Scarpetta *et al.,* 2000: 36ff.). However, changes in labour quality were of greater importance to labour input in Europe. The EU Commission's calculations show that the substitution of labour for capital affected labour productivity growth negatively in the EU in the 1995–9 period (–0.1 pp). This is attributable to the substantial employment creation Europe experienced in the last couple of years. Employment growth in the European Economic and Monetary Union, for example, increased from below one per cent in 1995 to roughly two per cent at the end of 2000 (Figure 17.7). While the reintegration of jobless persons in the labour market is certainly a positive development in the struggle to combat widespread unemployment, overall labour quality might actually have been negatively affected.

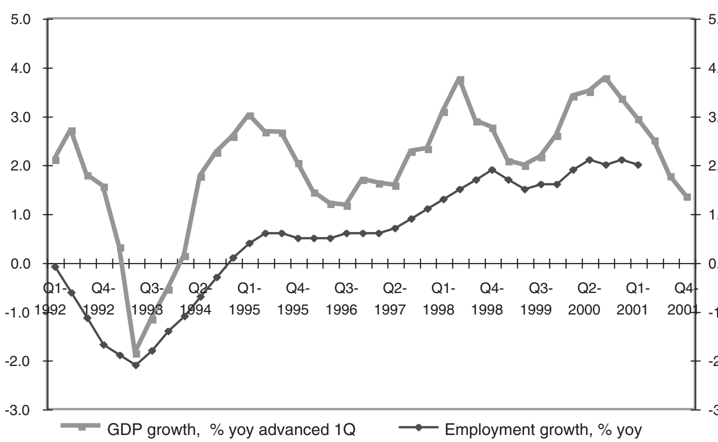


Figure 17.7 EMU: GDP and employment.

The OECD study finds that in some countries corrections for labour quality result in rising total factor productivity growth in the second half of the 1990s compared with the average of the 1990s, whereas without this correction TFP growth actually decreases. This is the case because a deterioration in labour quality leads to an underestimation of labour productivity and thus TFP growth. According to OECD data a deterioration of labour quality can be observed especially in Italy and France and, to a lesser extent, in Sweden. Presumably the rise in employment of less skilled workers in these countries in the second half of the 1990s affected TFP growth negatively.[14] Unfortunately, no data are available for Spain, Portugal or Greece where similar results could be expected. Evidence from labour market reform supports this hypothesis. A number of governments have stepped up their efforts to improve the employment situation in their respective countries, including France, Italy and the UK where in recent years a high number of unskilled workers entered the job market. As there is no consistent dataset for Europe with respect to labour quality changes a general assessment can hardly be made. However, in some countries decreasing labour quality certainly had adverse effects on TFP growth.

7 Conclusions

So far the 'new economy' in Europe is more mirage than reality. Even in the USA there are no clear signs of spillover effects from increased ICT investment on the efficiency of the economic process. In the USA, rising labour productivity growth was observed in the second half of the 1990s, whereas in Europe just the opposite was the case. While for Europe as a whole most studies show capital deepening with respect to ICT, evidence on TFP growth is rather disappointing. TFP growth actually declined during the second half of the 1990s. On an aggregated level, it becomes evident that – as in the USA – growth in labour productivity is most vigorous in the ICT-producing sectors, especially in manufacturing, and to a somewhat less pronounced degree in services. Spillover effects are hardly visible.

Using aggregated data might lead to misleading results, however. Europe is facing a north–south divide with respect to ICT penetration. Taking into account that a certain tipping point has to be surpassed for ICT to generate productivity-enhancing effects and that these will show up with a time lag only, it is justified to assume that northern Europe is more likely to experience new economy effects than southern Europe. However, labour productivity growth also decelerates on a country level. Total factor productivity growth, by contrast, seems to be in most cases higher, although not accelerating, for those countries which invest a large share of GDP in ICT. Moreover, a large ICT-producing sector seems to trigger strong TFP growth. Nevertheless, the observed slowdown in TFP growth for a number of countries cannot fully be explained. A major determinant might be the deterioration of labour quality due to the employment policies pursued in individual European countries over the past few years.

Having classified the Internet as a GPT it is likely that the full productivity-enhancing effects will be felt only with a significant time lag. The chances of spillover effects occurring – measured as rising TFP growth – are probably higher for the Nordic countries, the USA and Switzerland. In Italy and France, and probably also in Spain and Portugal, deteriorating labour quality as well as low levels of ICT penetration should prove counterproductive. New economy effects in these countries are likely to emerge farther down the road, if at all.

Notes

1 ICT is defined as information technology plus telecommunications equipment and telecommunications services.
2 Moreover, the new economy phenomenon goes hand in hand with other important changes in economic framework conditions, namely liberalization and deregulation, increased competition and more flexible labour markets. Also, globalization is often mentioned as an important driving force.
3 Labour productivity is measured as output per employee or output per hour worked.
4 Total factor productivity describes the relation between output and a set of input factors whereas partial productivity measures concentrate on the relation between one factor of production and output.
5 They are thus to be distinguished from incremental technological change, where progress is observable only in small steps.
6 It is clear, however, that ICT spending is a more suitable indicator to analyse the economic impact of ICT than PC penetration or other quantitative indicators (Daveri, 2000: 6) and should be used in regression analysis. However, Internet and PC penetration can serve as a valuable proxy for the quantitative diffusion of technology in an economy and the incurred network externalities.
7 Gros refers to Daveri (2000) 'La New Economy in Europa', IGIER Bocconi, mimeo.
8 The EU Commission acknowledges that there is a sizeable difference between the two concepts for the European Union whereas it is negligible for the USA (European Commission, 2000: 100).
9 See also Daveri (2000) for results on four large EU countries.
10 These countries have employed extensive labour market policy tools in recent years.
11 Employment growth likely plays a role in this context as Gust and Marquez measure labour productivity per employee.
12 This argument refers to hedonic price measures in the US as opposed to conventional price measurement in most European countries. The hedonic method allows the deflation of nominal variables by taking into account qualitative differences in goods. These differences in measurement have a substantial effect on productivity in an international comparison (Gräf, 2001).
13 A second aspect relates to a structural versus a cyclical improvement in productivity. Both the US and Europe experienced an improvement in their cyclical position during the 1990s. It could thus well be that the acceleration in productivity in the US is predominantly a cyclical phenomenon which is more pronounced that the improvement observed in Europe.
14 In the UK, deteriorating labour quality can be predominantly observed in the first half of the 1990s whereas the correction is less significant for the second half. Nevertheless, correcting for the effect in labour quality results in a significant pick-up in MFP growth during the 1990s. See also Euroframe (2000) on labour market policies in different European countries.

References

Daveri, F. (2000) 'Is growth an information technology story in Europe, too?', Working Paper no. 168, Università die Parma and IGIER.

EITO (2001) *European Information Technology Observatory*. Frankfurt.

Euroframe Group of Research Institutes (2000) *The Economic Situation of the European Union and the Outlook for 2001–2002*, Euroframe Report.

European Central Bank (2001) 'New technologies and productivity in the euro area'. *Monthly Bulletin*, July: 37–48.

European Commission (2000) *The EU Economy 2000 Review*, Manuscript no. 71, Brussels.

Gordon, R.J. (1999) *Has the 'New Economy' Rendered the Productivity Slowdown Obsolete?*. Northwestern University and NBER, revised version.

Gordon, R.J. (2000a) 'Does the "new economy" measure up to the great inventions of the past?' *Economic Perspectives* 14: 49–75.

Gordon, R.J. (2000b) *Was the Economy of the 1990s a New One?*. Northwestern University and NBER.

Gräf, B. (2001) 'New Economy: "Produktivitätswunder" in den USA nur ein statistisches Phänomen?' *Wirtschaftsinformatik* 43: 363–8.

Gros, D. *et al*. (2001) 'Testing the speed limit for Europe', 3rd Annual Report of the CEPS Macroeconomic Policy Group, Brussels.

Gust, C. and Marquez, J. (2000) 'Productivity developments abroad', *Federal Reserve Bulletin* 86: 665–81, Washington.

Mackintosh, I. (2001) 'Internet and highways are as chalk and cheese', *Financial Times* 13, August 2001.

Nordhaus, S. (2001) 'Productivity growth and the New Economy', NBER Working Paper 8096.

Oliner, St. D. and Sichel, D.E. (2000) 'The resurgence of growth in the late 1990s: is information technology the story?', Federal Reserve Board Finance and Economics Discussion Paper no. 2000–20, Washington.

Roeger, W. (2001) 'The contribution of information and communication technologies to growth in Europe and the US: a macroeconomic analysis', Economic Papers, no. 147, European Commission.

Scarpetta, S., Bassanini, A., Pilat, D. and Schreyer, P. (2000) 'Economic growth in the OECD area: recent trends at the aggregate and sectoral level', OECD Economics Working Paper, no. 248.

Schreyer, P. (2000) 'The contribution of information and communication technology to output growth: a study of the G7 countries', OECD Working Paper 2000/2.

Shiller, R. (2001) 'Just another superhighway', *Financial Times*, 1 August 2001.

Speyer, B. (2000) 'The Internet – a new general purpose technology?', *E-conomics,* no. 5, Deutsche Bank Research, Frankfurt.

18 Beyond the dot.coms*

Siegfried Utzig

Today, the meaning of the title 'beyond the dot.coms' is ambiguous. This was not the case as recently as a few months ago. If we focus first and foremost on the companies that use the Internet to provide their services – that is the dot.coms themselves – it is not only the development of their share prices that would lead us to assume we are dealing with a speculative bubble without any enduring effect on the real economy. After the frenzy of the last few years, companies providing services over the Internet are, in fact, facing a serious crisis. Just over a year ago, all signs were still pointing unequivocally to a continued boom and the new start-ups that were shooting up like mushrooms were finding it difficult to recruit staff. In summer 2001, however, the situation has changed totally. Since January 2000, at least 555 Internet companies have had to close down worldwide. Sixty per cent of these closures took place in the first six months of 2001.[1]

The impact of the closures on the labour market has been equally dramatic. In the USA alone, around 880 dot.coms have laid off more than 130,000 staff since the beginning of 2000.[2] This trend shows the enormous difficulties evidently faced by Internet companies in creating and implementing a profitable business model. But are we to conclude from this that the new economy does not exist? Does 'beyond the dot.coms' mean 'back to the old economy'?

Some remarks on the controversy over the new economy

This interpretation was naturally not intended, and it would, moreover, also be wrong. There can be no going back to the era before the Internet. The facts tell us that technology investments have plummeted over the last twelve months, first in the USA and then in Europe. There are essentially two explanations for this development. First, the impact of the specific laws of the net economy, such as the first-mover advantage, may have been underestimated or may have taken effect more quickly than anticipated. Second, expectations of the speed of the transition to the new economy were obviously wildly exaggerated. Hermann-Josef Lamberti, for example, a member of the board of Deutsche Bank, admitted that the fact that companies must first establish a network with one another before

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

B2B marketplaces can develop had been totally underestimated.[3] Neither consumers nor companies have changed their behaviour rapidly enough and for long enough to build up the kind of customer base which was expected in the short term and which is a prerequisite for profitability. The developments of the last twelve months are consequently a process of adjustment after the exaggerated expectations of the preceding years.

This has, without doubt, put a damper on the notion of consumers who do everything online. But whether the companies which failed were simply ahead of their time or whether they mistook the needs of a small group of consumers for the needs of all consumers has little relevance for the future of the new economy. It is the business models which failed, not the new economy itself. Neither the present crisis facing the dot.coms nor the slowdown in technology investments is going to reverse the process of global electronic networking. The issue of whether, and how, the new information and communications technologies will increase economic output and raise the standard of living therefore remains firmly on the agenda. 'Beyond the dot.coms' asks who the actual winners of the new economy are.

Up to now, whether or not you consider yourself one of the winners has seemed to depend on how you define the new economy and where you live. If you take a narrow definition, i.e. limited to the technology sector itself, and if you live in the USA, the answer will be a clear 'yes'. The narrow definition identifies the new economy with the dramatic acceleration in productivity tied, to an important degree, to innovations in information technology. Productivity refers to output per hour, and it is perhaps the single most important determinant of economic well-being, closely related to real income *per capita*.[4] If, on the other hand, you see the development of productivity in the economy as a whole as the key factor, or if you live in Europe, the reply will be much more cautious or will deny that the new economy has had any discernible effect.[5]

While the information and communications technologies themselves unquestionably made a significant contribution to economic growth during the 1990s, opinions differ about whether this increase in productivity has had any spillover effects onto the economy as a whole.

There are at least two competing explanations of the recent productivity data in the USA, only one of which contradicts the new economy hypothesis:

1. *The diffusion hypothesis*: This hypothesis says that several years or even decades have to pass before a new technology has diffused enough to significantly raise industrial productivity. But once the new technology has spread and people have learned how to use it, things really take off. The diffusion hypothesis supports the new economy hypothesis if interconnectivity has been the vehicle through which diffusion finally takes place.[6]
2. *The concentration hypothesis*: This hypothesis argues that the entire acceleration of total factor productivity growth – which Robert Gordon has estimated to be about 0.6 per cent per year once measurement changes and cyclical influences are accounted for – can be traced to the computer manufacturing

> industry. This one industry, while just a small piece of the overall economy, has produced a genuine productivity miracle since about 1995. If Gordon is correct, all the computer-using industries together, which comprise virtually the entire US economy, have displayed no net productivity acceleration at all. That appears to be bad news for the new economy hypothesis, which emphasizes the gains from using computers, not in manufacturing them.[7]

Are these findings enough to prove that the new economy hypothesis belongs in the realm of fantasy? Have we witnessed nothing of economic significance except for a temporary increase in productivity in a single industrial sector? Several macroeconomists seem to wish to give this impression. But the findings that have been put forward constitute the basis for a debate rather than put an end to it. This paper is not intended to be a contribution to the literature on TFP.[8] But the prominent position that this concept currently occupies in the new economy debate requires a few observations to be made on its suitability to reflect the changes which are associated with the new economy.

'Manna from heaven' or 'the measure of ignorance' are phrases commonly used to describe TFP. With a certain charm, they illustrate the central weakness of the concept, which nevertheless continues to enjoy widespread popularity. There are two aspects, in my view, that are significant in the context of the new economy debate.

First, there is the fact that US statistics use hedonic price indices to adjust for changes in quality in the computer industry. The strong growth measured by these indices is at the heart of the American productivity miracle. The direct consequence is an increase in real investment in information technology in the rest of the economy. Improvements in the quality of products and services outside the computer industry are not captured with the help of hedonic price indices, however. Little is known as yet about the extent to which the selective change in measuring improvements in quality leads to an understatement of the increase in productivity in the economic sectors outside the computer industry. It would seem reasonable, however, to assume that this is the case. The question therefore arises to what extent the selective use of hedonic price indices can distort measurements of productivity and growth.

A second objection concerns whether the neo-classical growth model which lies behind TFP is a suitable instrument to shed light on an economic period characterized not by improvements in familiar technologies, but by innovation, by imbalances and structural breaks. Much of the welfare gain from innovation comes from the production of better goods, and not just from the production of more goods. This is what a shift in the production function means. Innovation that results in better goods is not part of the TFP story.[9]

But the measurement problems in the definition of output which involve a quality dimension are not limited to product-oriented innovation. There are also problems with the definition of output which involve a quality dimension without reference to innovation. What are, exactly, the units of output in banking? How would one measure these units and account for differences in the quality of service that are characteristic of competition among banks? Unless the nature of the output can be defined precisely, it is impossible to determine its rate of growth or to answer questions about the impact of quality-enhancing innovations like those of communications technology.

Not to mention the fact that improvements in the banking system brought about by the new information and communications technologies make a major contribution to improving the allocation efficiency in the economy as a whole.

A large theoretical literature shows that banks can reduce the costs of acquiring information about firms and managers and lower the costs of conducting transactions. By providing more accurate information about production technologies and by exerting corporate control, better banks can enhance resource allocation and accelerate growth. Similarly, by facilitating risk management, improving the liquidity of assets available to savers, and reducing trading costs, banks can encourage investment in higher-return activities.[10]

According to the neo-classical growth theory, this enhanced efficiency should result in an increase in TFP. Does this mean the debate on assigning TFP growth to individual sectors and on possible spillover effects is the result of measuring problems or inadequate econometric models? Be that as it may, the fact that neither the statistics nor theoretical macroeconomic concepts are able to provide conclusive evidence of the development of a new economy cannot, under these circumstances, be taken as proof that the new economy does not exist. The findings merely illustrate the tremendous difficulties involved in accurately capturing not only improvements in the quality of products, but also new products themselves.

The challenge of explaining the influence of the worldwide electronic network on the global economic system is thus likely to lie well outside the calculation of TFP. Beyond the dot.coms, a comprehensive structural transformation is taking place, which was triggered by the new technologies, is changing the economic system at both national and global level, and, under certain conditions, is also enhancing productivity. There are currently many indications to suggest that despite the sketchy results provided by the research carried out so far, the so-called old economy will be the main beneficiary of this transformation. This assumption is based on the expectation that the new technologies will alter the size and structure of industries and companies, lead to massive shifts in relative prices and thus bring about fundamental changes in the supply of goods and services.

What makes the new economy so different?

Let us venture a comparison with radical economic changes in the past. Around two hundred years ago, the invention of the steam engine heralded the start of the industrial revolution. This new technology revolutionized the process of converting raw materials into products. At the same time, it triggered social and political changes which influence our lives even now. Information and communications technologies are today's steam engine. They are revolutionizing every process associated in the broadest possible sense with the exchange of information, ideas, and goods and services.

This makes information and the knowledge gleaned from this information the key production factors of the new economy. In the world of economic models, information is either complete or, if not, asymmetrically distributed. But it always remains an exogenous parameter. In the real world, away from all equilibrium models, the gaining of an information advantage and the erosion of this advantage by competition ensure economic dynamics. To all intents and purposes, information is the raw material for economic activity, so to speak. It is needed to discover market opportunities, find trading partners, control production processes and the exchange of goods, and monitor the quality of the supplied services. Information is the raw material for economic activity within companies, between companies and between companies and consumers.

The costs incurred in the exchange of goods and services by the search for business partners, by bargaining and finally by monitoring performance largely determine the structures of an economic system.[11] This immediately becomes apparent if we assume that these costs are virtually zero. Services which are now performed by a company's employees could then, in theory, also be handled by self-employed individuals dispersed over a wide geographic area and connected only by a communications network.

This 'atomized economic system' is purely hypothetical in nature. Yet it gives us some idea of the type of economic system which would evolve if, for example, the costs of interaction outlined above dropped sharply within a short period of time. The lines of demarcation between economic sectors would have to be redrawn, as would the organization and size of companies. The structure of economic sectors and of company organization would undergo radical transformation. Numerous new market opportunities would be created. The question of which services should be provided within a company and which via the market would have to be totally reconsidered. It does not take much imagination to realize that the global communications network which is currently evolving is pushing things in just this direction. What is more, in some sectors, such as banking, for example, these changes have already begun to take effect.[12]

Which sectors will profit from the global network?

By reducing transaction costs

How will these changes triggered by the Internet manifest themselves in individual sectors of the economy? The most important quality of the global electronic network is also the most obvious. Information can be disseminated swiftly, conveniently and cheaply. It will be possible to carry out routine tasks, above all, at a fraction of the present cost. Transaction cost savings will therefore be particularly great in companies where output can easily be digitized and standardized.

This certainly applies to a large number of services associated with the new information and communications technologies. It also goes for financial services providers, including banks. But there are also other important areas in which global interconnectivity will not only reduce costs, but will doubtless give rise to further innovations too. The health service and public administration will probably be among those to benefit most from the electronic network. Both may be classified as data intensive. Enormous savings can be achieved by digitizing this data and by using the Internet. Yet it will be difficult to measure improvements in productivity in these sectors in the absence of market prices to use as a yardstick.

In the health service, for example, considerable savings could be made by digitizing doctors' and clinics' invoicing to the health insurance organizations. The potential savings here are quite comparable with those in the banking industry resulting from the transition to paperless payment transactions. If digitalization and communication over the Internet were also exploited for creating, updating and transferring patients' files, we could expect not just further cost savings, but a considerable improvement in the quality and efficiency of the health service as a whole. This presupposes, however, that a satisfactory solution can be found to the problem of data security and data protection, particularly from the patients' point of view.

The potential of the new economy at all levels of public administration has been largely ignored up to now. Resources are constantly tied up, for example, to answer queries about public services, entitlements to benefits and grants from public funds, etc. The Internet is an excellent means of disseminating such information among citizens at a fraction of present costs. The digital town hall, where citizens no longer have to face long waits but can find the various services supplied by individual government departments a mere mouse click away from one another, is likely to become a reality when digital signatures are recognized in the next few years. This does not only constitute a potential for greater productivity in public administration in the form of enhanced efficiency and improved services. Citizens and companies will also spend considerably less time travelling to and waiting around in public offices. The effects of the electronic network are therefore perfectly comparable to those of deregulation.

The same is true of the tax office. It is already possible throughout Germany to submit income tax returns via the Internet, though few people make use of this opportunity. For some other types of tax, such as trade tax and VAT, a few federal states also already accept digital tax returns.

In the financial services industry, the changes brought about by the Internet are first and foremost affecting retail banking. The traditional relationship between customer and bank, the result both of geographical proximity and the bank's familiarity with the needs of its customers, is now being challenged by the inexpensive processing of standard banking services over the Internet. These include queries about balances, domestic and cross-border remittances, buying and selling securities (direct brokerage), arranging, changing and cancelling standing orders.[13]

By more efficient management

The effects on the majority of sectors in the so-called old economy will be less direct, though just as far-reaching. The importance of the Internet as a management instrument is likely to continue gaining ground. In large companies, this process is already well underway. But it will also have an impact on small and medium-sized companies and trigger significant restructuring in all sectors of the economy.

Many of the anticipated efficiency gains will result from making use of the Internet for supply chain management. The Internet will enable just-in-time production, which has long been an established trend anyway, to be even more finely honed. In many companies, the anticipated efficiency gains and savings can already be seen in the area of procurement. The Internet will also enable another trend already familiar in the old economy to become even more firmly established – outsourcing. Strategic decisions by companies to concentrate on their core competences may be observed in all sectors of the economy, including the financial sector.

The changes will probably be felt most profoundly in industry, however. The meaning of the term manufacturing industry will undergo a radical transformation. In the coming years, we will see a sharp rise in the number of manufacturing enterprises which outsource more and more sections of their production to other companies while confining themselves increasingly to their core competence and to using the Internet for the necessary coordination between their suppliers. Yet focusing on entrepreneurial core competences is nothing more than an enhanced utilization of entrepreneurial know-how. Productivity effects are not to be expected in the sense of a movement on the production curve, but in the sense of a shift of the production curve – an increase in TFP.

The possibility to use the Internet to connect the staff of a company, or of several associated companies, with one another over large distances at little cost and thus to optimize cooperation is likely to result in further increases in efficiency. An example of this trend may be seen in the automobile industry, where enhanced productivity has been achieved in this way in product development, in the supply of components and in production. It would now be perfectly feasible for this industry to implement the production model perfected by the computer manufacturers Dell: the customer would put together the car he wants on the Internet, order it – and only then would it go into production. This would without doubt have severe implications for existing distribution structures.

By keener competition

This brings us to another consequence of the Internet – more intense competition. For both suppliers and customers, the Internet increases market transparency at a stroke for goods and services which can be precisely defined. Buyers can explore a large geographical area quickly and inexpensively to find the lowest price, while sellers gain access to a far larger pool of potential customers. Thanks to the Internet, actual market activity is moving closer towards the textbook model of perfect competition. In practical terms, the consequences are lower profit margins, more efficient production and greater consumer satisfaction. Ultimately, keener competition will also enhance price stability.

In the long term, the Internet will thus act as a great ‘leveller’ by helping to lower worldwide the barriers to entering the market. For a company anywhere in the world which has developed a better product, this will improve its chances of actually being able to sell the product, and thus force other companies to improve their own products. The price transparency and the geographical expansion of the market which can be achieved by the Internet will therefore act as a global stimulus for continuously increasing efficiency.

The success of the new economy is not automatically assured

At present, it is not possible to say which places in the economy will actually be able to gain from the new economy, nor predict the size of these gains. The above observations are intended to illustrate that the new technologies will influence the productivity of the old economy in very diverse ways. One effect will certainly be capital deepening. We can also anticipate enhanced efficiency, which will result in a one-off rise in the level of prosperity. We will also see an increase in TFP.

But the comprehensive structural transformation should not lead us to make the mistake of judging the success of the new economy by the yardstick of productivity growth alone. Keener competition and falling profit margins are not synonymous with an increase in productivity. In many cases, the drop in prices will probably merely be passed on. Ultimately, the consumer alone will benefit. This process, though on the whole desirable, does not mean an increase in productivity. In many other sectors too, such as online trading, the most obvious consequences of the new economy will be more choice, diversity and convenience. It will be difficult to capture these qualitative improvements in a growth of GDP.

The extent of the increase in prosperity will, however, also largely depend on how quickly the use of the Internet spreads in companies and, above all, among the population at large. There is still a clear discrepancy in the use of the new technologies between companies and private households. In medium-sized and large-scale enterprises, use of the new technologies has probably become the norm. Small companies will have to follow suit to remain competitive. Private households, on the other hand, still have considerable ground to make up.[14]

The statistics show quite clearly that use of the new technologies in private households is still very much dependent on education and income. There is already much talk in some circles of a 'digital divide', meaning that people on a low income are excluded from using computers and the Internet. Political steps can be taken to reverse this development by installing Internet connections in schools, for example. The steady and rapid fall in prices for Internet and mobile phone use will also help to bridge this divide. All in all, this is likely to ensure that every citizen will have the opportunity to reap the benefits of the new economy.

Whether or not the new economy will succeed in developing its full potential also depends on a number of other factors, however. One important issue is whether the new technologies will really be used in the individual sectors of the economy to the extent currently anticipated. And, above all, whether the associated changes in company practices and business models really take place. A look at investment in information technology in the past shows that the expected increases in productivity often failed to materialize. These cannot be prevented in the long run, because the companies whose IT investment proves successful will gain a competitive edge. But management errors, or resistance and inertia in existing organizational structures might slow down the process. We lack the empirical values on which to base a reliable forecast of the timescale and extent of these changes in productivity.

The success of the new economy can only be assured by economic policy

All entrepreneurial endeavours to successfully exploit the new technologies' potential to increase productivity are doomed to failure, however, if the structural changes which are initiated are not supported by the general economic environment. An economic policy geared towards conserving existing structures will tend to obstruct the transition to the new economy and the anticipated acceleration in the rate of growth. In this context too, the findings of the ECB survey hold no surprises.[15]

The turnaround in the US productivity trend and the transition to the new economy should therefore not be viewed in isolation. Both can be interpreted as the peak, for the time being, of a development whose origin is to be found in the shift in paradigms of American economic policy in the early 1980s towards supply-side policies. The development of the US economy since then suggests that this change of course has been largely responsible for much of the success. The continuous improvement of conditions conducive to growth also explains much of the success in realizing US growth potential in the 1990s. Perhaps the sustained prosperity phase was even the prerequisite for the development which is known today as the new economy.

No changes in macroeconomic policy areas

Consequently, the new economy does not require a new economic policy. The traditional 'old' macroeconomic rules will continue to apply and form the basis of economic policy in the 'new' era too. The goals of monetary policy and fiscal policy will remain essentially unchanged. A high level of monetary stability, along with budgetary consolidation and low tax rates are still major prerequisites for stable economic growth. However, the following questions must be asked in both policy areas:

- Has the economy's sustainable growth rate been raised? And might it go even higher?
- Has the lowest unemployment rate consistent with stable inflation – the so-called NAIRU (Non-Accelerating Inflation Rate of Unemployment) – been reduced? And if so, how far?

The current debate shows that it is not always easy to recognize a change in the long-term growth rate, and if it is recognized, then only after a considerable time lag.

For years after the 1973 productivity slowdown, for example, economists, policymakers, business executives, and workers continued to base decisions on what we now realize was an excessively optimistic estimate of the productivity trend. As one concrete example of the large errors that can be made when perceptions lag behind reality, it has been argued that the Federal Reserve pursued a highly inflationary monetary policy in the 1970s largely because it overestimated the productivity trend.

What might happen if the productivity growth trend falls, but neither labour nor management recognizes that for a while? If wage agreements continue to be based on an assumed productivity trend of, say, three per cent, but actual productivity growth falls to one per cent, then unit labour costs will rise two per cent faster than either labour or management expected. The overestimate of productivity growth will lead first to excessive real wage agreements (relative to productivity) and then to higher costs and inflation. Firms will react by reducing employment and raising prices at the same time. This scenario helps explain the stagflation of the 1970s.

Now turn the logic around. If productivity growth speeds up, but people do not know it, wage agreements will come in too low (relative to productivity), thereby lowering business costs. Employment should also expand since labour gets cheaper (again, relative to productivity). This scenario may explain part of the fabulous macroeconomic performance of the late 1990s: an unrecognized productivity acceleration led to both higher employment and lower inflation.

It is a bit harder to understand why improvements in information technology should reduce the NAIRU permanently. Information technology might reduce the frictions involved in matching workers to jobs. More generally, advances in information technology make it possible to shift some economic activities to locations where labour is more readily available. But it is hard to believe that such effects could be quantitatively large on an economy-wide basis.

Consider, however, two believable scenarios that might produce a transitory decline in the NAIRU. One is the misperception scenario just described. If actual productivity grows faster than perceived productivity, the economy will experience a surprisingly favourable combination of stable inflation and low unemployment. In the data, it will appear as if the NAIRU has declined. But as perceptions adjust to the new, faster pace of productivity gains, the apparent NAIRU should return to normal.

This analysis points to the following tentative conclusions for monetary policymakers:

- The current revolution in information technology may indeed have accelerated the nation's productivity growth. If so, central banks should allow the economy to grow faster than it would with a lower productivity trend.
- An unrecognized acceleration of productivity may reduce the NAIRU, but only temporarily. If so, the apparent NAIRU will return to a more normal value as perceptions catch on to the new reality.

Yet whether or not the success of the new economy in the USA can be repeated in Europe does not essentially depend on monetary and fiscal policy. There are no longer any significant differences between Europe and the USA as far as the orientation of their macroeconomic policy is concerned. Both have a stable macroeconomic environment. But the difference in their economic development shows that sound monetary and fiscal policies alone are evidently not enough to generate the increase in the innovation potential which has made the USA a benchmark for the new economy. The success of the new economy in Europe will therefore be largely determined in policy areas which can influence the flexibility and adaptability of market participants. In concluding that there are, as yet, no signs of a new economy in Europe, the European Central Bank is implicitly highlighting the fact that there is still a need for economic policymakers to take action. If a course is to be set for the new economy, a series of individual economic policy measures must be implemented, above all in areas where there have long been calls for reform in Europe.

Confidence in market forces

Successful structural change and dynamic economic growth are closely connected with the opportunities to allow entrepreneurial potential to develop freely. The extent and intensity of entrepreneurial activity depend, in turn, on a number of preconditions. These range from free market access and corporate financing opportunities to the standing enjoyed by an entrepreneur in public opinion. Significant differences may be observed between the economic policies of Europe and the USA in the extent of their confidence in the efficiency of market forces.

The policy of deregulation, which made a major contribution to the US economic upswing in the 1990s, serves as a good example. America began to deregulate telecommunications in the early 1980s. The competition which this unleashed resulted in the prices of telecommunications services dropping sharply. This increased demand and created incentives for new, innovative services. The concurrence of this development with the decisive technological progress in communications technology provided the basis in the years that followed for the above-average increase in productivity and high employment in the USA.

In Europe, on the other hand, the deregulation of the markets for telecommunications services lagged behind by a good ten years. For this reason, there was also a considerable time-lag before we saw the beginnings of competition in products and prices. In the meantime, market conditions in Europe and the USA have more or less converged in the area of telecommunications. Nevertheless, existing obstacles to competition still have to be dismantled in local networks to make it less expensive for consumers and companies to use the Internet.

The example of the deregulation in telecommunications and the flourishing of communications technology companies clearly illustrates that economic policy can only be successful if it is farsighted enough to create entrepreneurial room to manoeuvre and does not attempt to protect existing economic structures. The experience of the past ten years has proved beyond a doubt that the jobs of the future will be created, above all, in companies which do not yet exist today, or which have only existed for a short time. Whether, and if so how rapidly, structural change can gain momentum and create jobs depends not least on how quickly these new companies can assert themselves in competition with the companies already on the market. The staying power of 'old' companies and industries often depends not only on their competitive market position, but also on their ability to bring their interest in conserving existing market structures to bear in the political arena.

In the past, the staying power of 'old' companies was evidently far greater in Germany and the rest of Europe than in the USA. It is interesting to note that in the last forty years, no business start-up has managed to work its way into the league table of Germany's largest enterprises. The twenty-five largest German companies in 1999 were already in the top twenty-five in 1960. The situation is totally different in the USA. One in three of today's largest twenty-five companies did not exist at all forty years ago. These 'newcomers' include a number of IT companies with a commanding competitive position on the world market.

The general conditions facing business start-ups are also key to successful structural change. The attraction of running one's own business in the USA and the availability of capital to finance new businesses without doubt contributed greatly to the boom in new start-ups in the 1990s. In Germany and other European countries, there has been some improvement in the last few years in the conditions for setting up a business. There have been numerous start-ups in the new growth industries of information and communications technologies, including the associated service sectors, and in the area of biotechnology as well. This trend gives cause for hope that entrepreneurship is about to be rediscovered in Europe.

The labour market must become more flexible

Structural change – particularly in a form as radical as that currently underway – always has a concrete impact on the labour market, since the declining number of jobs in one sector very rarely coincides in time or place with the new employment opportunities to emerge in another. The discrepancy in professional qualifications is likely to be even more important, however. If the economic transformation is not to lead to an – albeit temporary – increase in unemployment, a high degree of flexibility is required on the labour market. Creating this flexibility must be labour policy's top priority in the new economy.

But a flexible labour market is also necessary for another reason, and one which is often overlooked. The new economy does not only require highly skilled workers, it will also lead to different working conditions. Lifelong employment with its associated social security benefits, which characterized the industrial society, will gradually give way to a new trend towards temporary working contracts, part-time employment and various forms of self-employment. Different employment options will be exercised simultaneously or one after the other, perhaps interspersed with phases without employment.

The USA is without doubt closer to the ideal of a flexible labour market than the countries of the EU. In the USA, there are no industry-wide wage negotiations between employers and trade unions, which in Germany keep the cost of labour too high, cement existing wage structures and are responsible for a collective shortening of the working week. In 1998, the pay agreements of only around ten per cent of employees in the US private sector were negotiated collectively. In light of the fact that wage negotiations are more or less unregulated, it is unsurprising that unlike in Europe, US economic forecasts take little account of wages policies.

Another characteristic feature of the American labour market is the traditionally wide wage spread. Furthermore, the USA has succeeded in structuring social transfer payments in such a way that there is still an incentive to take on even low-paid employment. The American labour market is thus well placed to absorb the impact on employment of a macroeconomic structural change.

In contrast, regulation of the labour market in Europe is diverse and complex, and badly placed to meet the challenges posed by the upcoming changes. The lack of flexibility on wages and working hours, and the sometimes still extensive regulation of the labour market might jeopardize the general public's acceptance of these changes since they make it less likely that the new economy's success in creating jobs in the USA can be repeated. A fundamental reform of the labour market, which would allow greater flexibility on wages, working hours and types of employment, is therefore one of the most pressing prerequisites in Europe for the new economy.

There is as yet no evidence of progress in this direction, at least not in the biggest economies of the euro zone. On the contrary, there have even been some backward steps, such as the introduction of the 35-hour week in France. Nor can the amendment of the Labour–Management Relations Act in Germany to prevent the termination of collective pay agreements be regarded as progress in the sense of a flexible labour market.

Social security in the new economy

The social security systems in almost all industrialized countries have been in considerable need of reform for some time. This is closely linked with demographic developments in these countries and therefore arose quite independently of the structural changes associated with the new economy. Pensions and the health system are without doubt in most urgent need of reform. Progress differs from one country to another. But it is clear that Europe will continue to build on its socio-political tradition and that the US will therefore be unable serve as a model in this area.

There is a consensus that the reforms of the social security systems must result in long-term solutions which can guarantee citizens a secure and reliable income in old age, and cover health risks and the risk of unemployment. But this can only be achieved if social security reforms take due account of the massive changes triggered by the new economy in economic processes and in employment. Generally speaking, however, this has not been the case so far.

In Germany, for example, the state pension insurance scheme has been based up to now on the assumption that a person eligible for a pension will normally have been in lifelong employment and paid regular social security contributions. The pension reform which has now been adopted continues to be based on such a scenario. The changes on the labour market which have been sparked by information and communications technologies, however, mean that this form of employment, which has its roots in the industrial age, will steadily decline in importance in the coming years. If we persist with the usual pay-as-you-go system based on this traditional form of employment, there will be rising numbers of those without adequate cover, or without any cover at all, and the burden on the remaining contributors to the system will become ever heavier. Thus if we ignore the changes which the new economy has triggered in the labour market, we will achieve the exact opposite of the original objective. A positive development, on the other hand, is the fact that there is now broad realization that the state pension scheme must be supplemented with private and occupational funded schemes.

Conclusions

The availability of the new technologies and inexpensive access for broad sections of the population are not sufficient in themselves to ensure that their enormous economic potential will be developed to the full. The new economy has its own laws and requires more flexibility from market participants. It can therefore only flourish if general national and international conditions allow these requirements to be fulfilled. For this reason, the new economy is more than the reflection of a technological revolution. It highlights the responsibility of economic policymakers to create a framework within which market forces may freely interact and develop. Policies based on the concept of '*Ordnungspolitik*', which are often set aside on the grounds of ostensible macroeconomic necessities, are likely, under these circumstances, to come back into their own. It is no easy undertaking, however, to improve the general conditions for innovation and entrepreneurship so that they can meet the challenges posed by the technological transition to the information society. Resistance to such reforms is considerable, especially in Europe.

Reforms are rendered even more difficult by the experience of the USA, which shows that there are costs associated with the new economy as well as benefits. An acceleration of economic growth, with the associated revolutionary changes in the structure of the economy, does not take place without friction. Both existing and newly created real and human capital will become obsolete more rapidly. This increases the burden of adjustment both for companies and for each and every individual. The new economy means more opportunities, but more risks too.

Nevertheless, there is no reason to be excessively pessimistic. In Europe, the radical changes that may lead us beyond the dot.coms are already underway in many areas. Yet there are still barriers holding us back from a fundamental breakthrough. The dismantling of obstructive market regulations is prevented again and again by forces which are still firmly clinging to the traditional system. This is slowing down competition and growth. The success of the new order is being impeded. Europe's new economy will only be able to develop its full potential when harking back to the industrial society and the overemphasis on the risks associated with change are replaced by more openness to the technical innovations of entrepreneurial activity.

Notes

1 Cf. webmergers.com, mid-year report.
2 Source: The Industry Standard.
3 Lamberti, H.-J. (2001) Interview in *Die Welt*, 16 July 2001.
4 Meyer, L.H. (2001) 'What happened to the New Economy?', Remarks before the New York Association for Business Economics and Downtown Economists, New York, 6 June 2001.
5 Cf. Gordon (1999) 'Has the 'New Economy' rendered the productivity slowdown obsolete?', NBER Working Paper, ECB (2001). 'New technologies and productivity in the euro area', July 2001 monthly bulletin, pp. 37–48.
6 Paul David argued with the example of the electric dynamo that such diffusion delays might explain the computer paradox. See David, P. (1990) 'The dynamo and the computer: An historical perspective on the modern productivity paradox', *American Economic Review*, 80, Papers and Proceedings, pp. 355–61.
7 Gordon, R. J. (2001) 'Technology and economic performance in the American economy', Discussion Paper, July 2001.
8 Cf. for example: Hulten, C.R. (2000) 'Total factor productivity: A short biography', NBER Working Paper 7471.
9 Hulten, C.R. (2000) p. 40.
10 Beck,T., Levine, R. and Loayza, N. (1999) 'Finance and the sources of growth', World Bank Working Paper 2057.
11 Cf. Coase, R.H. (1937) 'The nature of the firm', in *Economica* 4, pp. 386–405.
12 Cf., for example, Picot, A. and Neuburger, R. (2000) *Banken und das Firmenkundengeschäft im Internet-Zeitalter*, Cologne.
13 Cf. Bundesverband deutscher Banken (2000) *Banking Survey 2000 – the challenge of the information society*, Berlin
14 Cf. BITCOM (2001) *Wege in die Informationsgesellschaft*, Berlin/Frankfurt.
15 ECB (2001).

19 International productivity differences

Explanations and policy implications*

Johan Van Gompel

Abstract

The trend in and the level of productivity in a country are subject to the influence of a complex set of interactive variables of all kinds. Attempts to isolate the importance of the determinants are generally not easy because the relationship with productivity is often not one-way or because productivity (especially in the short term) is carried along on the waves of the economic cycle. In recent years, the role of information and communications technology (ICT) in promoting the strong productivity dynamics in the USA after 1995 has received particular attention in economic literature. Although the scale of the impact of ICT is not certain and there is also a great deal of discussion still about the spillover effects for the economy as a whole, the majority of economists are now assuming that ICT offers an (or at least some) explanation for international productivity differences. In this article, we discuss the extent to which the utilization of technological progress in the production process is influenced by the structural–institutional characteristics of a country, and which policy measures the government can take (or must avoid) to allow innovating economic activities to develop to the full. The characteristics alluded to are varied and form the framework in which people develop skills and companies innovate and invest. In particular, it is a matter, for example, of the working of the labour market, costs of adjustment with regard to the capital stock, regulation, efforts in the field of research and development (R&D), education and training and the degree of (international) competition.

1 Productivity: importance and determinants

Productivity reflects the extent to which companies succeed in deploying the production factors used (mainly labour and capital) efficiently in the production process. Economists usually measure it as the ratio between the added value of the goods and services produced, mostly in constant prices, and the factors of production used for this. The inclusion of several factors at once in the productivity standard allows substitution effects between the factors to be brought into account and is therefore, from the theoretical point of view, the most suitable

* Paper presented at 23rd SUERF Colloquium, Brussels, 27 October 2001.

way of measuring the overall efficiency of the economic process. It is then a question of the so-called ‘total or multi-factor productivity’. In practice, it is not easy to measure, however (see Appendix 19A), and for this reason productivity is usually approached from a single factor (so-called ‘partial productivity’). This is usually labour, since, compared to capital productivity, labour productivity is more easily measured. The capital input generally consists of highly diverse goods (machinery, buildings, computers, etc.) and it is not always clear which part the production process ‘uses up’ in a given period. Labour productivity can be measured on the basis of the number of workers or via the number of hours worked by the workers. The latter method of calculation is preferred, since it takes account of overtime worked and the trend towards the shortening of working hours and part-time work.

The trend in productivity is of both social and purely economic importance. On the one hand, positive growth in labour productivity is usually associated with growing material prosperity of the population. The higher real production which derives from a more complete and more optimal deployment of workers allows higher real wages and consequently a broader spending pattern. The importance of the productivity growth figure appears from the time needed to double the production per unit of labour deployed. Where the annual increase in labour productivity is 2 per cent, for example, the time needed is 35 years, where growth is only 0.5 per cent it will take no less than 140 years. An increase in labour productivity is only unequivocally positive for society as a whole if it does not lead to exclusion of labour. If the proportion of employed persons in the total population declines, labour productivity and real income *per capita* move in opposite directions. Naturally, differences between the GDP per worker and the GDP *per capita* also have to do with shifts in demographic factors, such as the ageing of the population (see also Figure 19.1). In contrast to labour productivity, capital productivity is of less importance for prosperity. A country may have high capital productivity and yet be very poor. It is possible that underdeveloped countries use the little capital at their disposal in such a way that it has quite high productivity per unit invested. But that does not mean that there is a high level of prosperity there. The labour productivity will be low there and then a high capital productivity is of little avail. High capital productivity on the other hand makes investment attractive and the larger supply of capital goods deriving from this in turn influences the labour productivity, as a result of which there may nevertheless be an indirect effect on prosperity. In addition to their importance for prosperity, international differences in productivity trend on the other hand also indicate whether a country is gaining or losing competitiveness in relative terms, although the relative costs and prices at which goods are manufactured and sold also play a role here. In any case, important interrelationships exist between these nominal variables and productivity. For instance, productivity may rise sharply through an accelerated substitution of labour for capital as a result of excessively high wage rises. In the case of this so-called ‘perverse’ rise in labour productivity, the normal relationship between the trend in productivity and wages is reversed. For countries

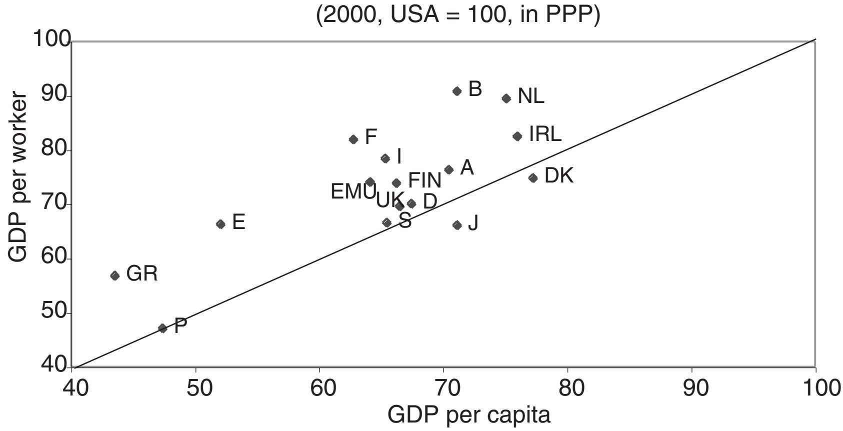


Figure 19.1 Prosperity and productivity.
Sources: European Commission, OECD.

where a relatively strong increase in labour productivity is accompanied by a sharp contraction in employment, the use of wage costs per unit of production is consequently less suitable as an indicator of competitiveness.

The trend in and level of productivity in a country are subject to the influence of a complex set of interactive variables of all kinds. Attempts to isolate the importance of the determinants are generally not easy because the relationship with productivity is often not one way or because productivity (especially in the short term) is carried along on the waves of the economic cycle. Short-term fluctuations in productivity are usually procyclical, which means that productivity increases at times of economic boom and declines at times of economic slump. This occurs because there are costs associated with the recruitment, training or dismissal of workers and with the expansion or contraction of the supply of capital goods (costs relating to feasibility studies, implementation, etc.), as a result of which companies do not carry out definitive changes immediately. So long as companies are not certain that it is not merely a chance or temporary surge or decline in economic activity, they will try to get to grips with the changed situation flexibly. This is possible by deploying temporarily superfluous workers in maintenance tasks or by adapting the working hours of their workers and the capacity utilization rate of their machinery. Research has shown that demand shocks related to the economic cycle, such as changes in interest rates, money supply or government spending, can explain up to about a quarter of the short-term fluctuations in productivity (see e.g. Evans, 1992).

In the longer term, the economic trend can also generate permanent productivity gains by purging inefficient companies or dismissing less productive workers during recessions (‘cleansing effect’), or by incorporating the latest techniques more rapidly into the capital stock during longer periods of sustained high economic growth (see Atella and Quintieri, 1998). The rate of technological progress is for that matter of a more fundamental nature for the long-term trend in productivity. Here it is a matter of qualitative improvement of the factors of production, as expressed in investments in new machinery or production techniques and processes, in the better education of the workers or in a more efficient company management and organization (as regards marketing, method of financing, commercial dynamism, etc.). Companies may increase these so-called 'deepening' in physical and human capital both through their own efforts in the field of research and development ('performed R&D') and through the dissemination and imitation of innovations brought about elsewhere ('acquired R&D').

2 Productivity paradox

Compared to the post-war decades, the industrialized countries experienced a significant slowdown in average annual growth of macroeconomic productivity in the period after 1973 (see Figure 19.2). The average annual growth of total factor productivity in the private sector fell in the OECD by 1.7 per cent in 1973–95, compared to 1960–73 and that of labour productivity by as much as 3.3 per cent. This decline came as a surprise, in view of the worldwide massive increase in investment in the application of new technologies (mainly in the field of electronics, computers, telecommunications and biotechnology) against a background of greatly increased international interrelationships. This contradiction is known as the 'Solow paradox' for which a large number of explanations are now available in the literature (see e.g. OECD, 1991). A first is that the reference period (i.e. the quarter of a century after the Second World War) is not very representative. In this period,

Figure 19.2 Productivity trend (private sector, average annual increase, in %).
Source: Thomson Financial Datastream.

the industrial infrastructure in Europe and Japan had to be largely rebuilt, which offered unique opportunities to incorporate modern technologies, mainly from the USA, into the production machinery. As a result of these technology transfers, the industrialized countries were able to make use of a considerable growth potential in the post-war decades and as a result start to catch up with the USA.[1] The decline in the impact of technology on the development of productivity in these countries after 1973 is claimed then to reflect the process of diminishing returns from imitation, as a result of the potential to catch up, drying up.

Another explanation for the paradox is that, together with diminishing returns in the technology-imitating countries, declining efficiency of technological innovation occurred in the leading technology countries themselves. This was apparently attributable, *inter alia*, to a growing tightening of production structures as technologies were adjusted, which had a negative impact on the possibility of introducing further important innovations. A 'technology trap' of this kind lasts until a new set of basic innovations has sufficient power to make a breakthrough and to bring about a fundamental change to the existing technological paradigm.[2] This most probably partly explains why, in spite of the IT boom, computers worldwide until about ten years ago all in all still only constituted a small fraction of the total stock of business assets. Just as with, for example, the dynamo, the transistor and the electric motor, it took some time before new computer technologies were fully disseminated, which can be explained *inter alia* by sunk investment costs in the existing production infrastructure, teething troubles, the lack of complementary technologies, such as software, or by the time needed for training.[3] Moreover, immediately after their launch, new technologies usually led simply to user-friendliness, substitution for existing products or to a redistribution of market shares in favour of more efficient companies, and therefore was not always (entirely) expressed in the measurement of macroeconomic productivity.

Still further explanations for the paradox are based on the finding that the worldwide slowdown in productivity went hand in hand with major structural changes and with changed government intervention. As regards structural changes, it is a matter *inter alia* of the improvement of working conditions or the increased participation of women on the labour market. The changes alluded to as regards government policy relate to, for example, the decline in government investment, higher tax rates, stricter environmental regulations, the development of a social security system, etc. These changes, which it is claimed led to a more unfavourable business climate, were also accompanied by successive shock movements, such as the oil crises in the 1970s and the accompanying hefty wage rises, the scale of the exchange rate volatility after the definitive end to the Bretton-Woods system in 1973 and the advent of the emerging markets in Asia and elsewhere. These were quite rapidly disseminated internationally through the increased interrelationships between the industrialized countries and brought weaknesses, such as lack of labour market flexibility, to the surface.

3 Importance of ICT

Although the paradox was more or less general, in the past thirty years substantial international differences in productivity trend in the private sector remained. A striking fact was the relative position of the USA in relation to the other industrialized countries. In a number of European core countries (Belgium, The Netherlands, Germany, France and Italy), the slower growth in labour productivity continued even after the economically turbulent 1970s. In countries which experienced little or no growth in labour productivity during the 1970s, such as the UK and the USA among others, this growth picked up again after the oil crises. The strong recovery in productivity in the USA after 1995 was especially striking. The average annual growth in labour productivity there shot up to 3 per cent, compared to only 1.2 per cent in the period 1973–95. In the EMU and Japan, on the other hand, the annual increase in labour productivity in the second half of the 1990s fell by around 1 per cent, compared with 1973–95 (see Figure 19.2). Nevertheless, the accumulated growth in the American labour productivity over the past thirty years remained 18 percentage points below that in the EMU and 27 percentage points lower than the Japanese.

As a result of this relatively unfavourable trend in the USA up to 1995, a number of European countries (including Belgium, The Netherlands and Norway) were nearly able to keep pace with the USA in terms of productivity level. Other European countries too made a leap forwards, larger in some cases than in others, but in 2000 still had labour productivity which was between 20 per cent to even more than 50 per cent lower than the American level. The greatest movement to catch up with the USA since 1960 was achieved by Ireland, whilst Sweden, the UK and Germany improved only slightly compared to the American level during this period (see Figure 19.3). The movement to catch up was greater

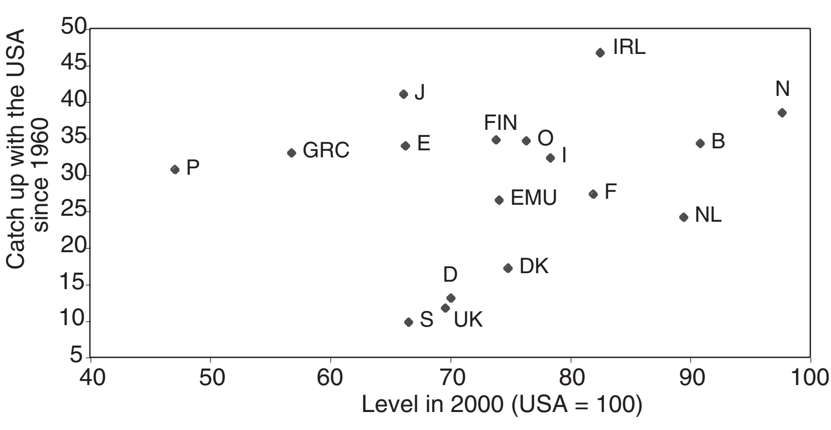


Figure 19.3 International differences in labour productivity (total economy, per worker, in PPP).

Source: Own calculations based on European Commission.

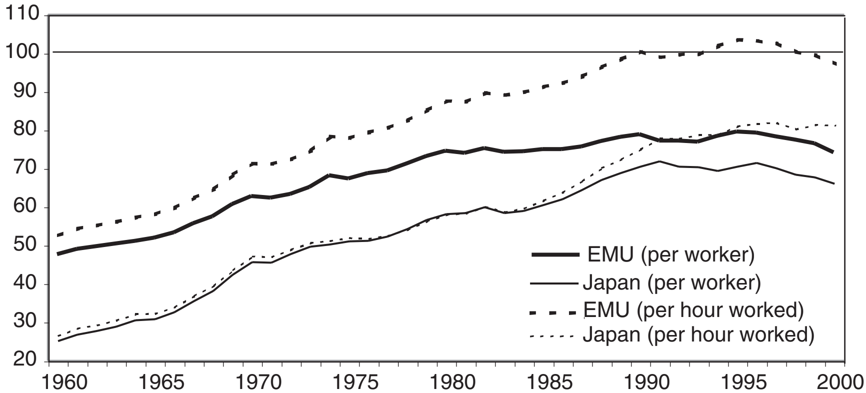


Figure 19.4 Relative level of labour productivity (total economy, in PPP, USA = 100).
Source: Own calculations based on European Commission and OECD.

in Japan than that of the EMU as a whole, but the Japanese productivity level in 2000 remained a good third below that of the USA, compared to about a quarter for the EMU. The movement by the industrialized countries to catch up with the USA was usually greater for productivity per hour worked than for that per employee. This indicates a wider gap with regard to the annual average number of hours worked per person. This figure has declined by 8 per cent in Europe and 15 per cent in Japan since the beginning of the 1980s, partly as a result of measures to reduce working hours, whilst it has increased by 5 per cent in the USA (see Figure 19.4).

To date, economists have not reached agreement about the origin of the relatively rapid growth in productivity in the USA after 1995. One point of discussion is whether the productivity improvement observed is largely a cyclical phenomenon or is of a more lasting nature. According to proponents of the 'New Economy' paradigm, the productivity growth trend has come to lie definitively at a higher level under the impact of new information and communications technologies (ICT). In line with this, the calculations of the American Council of Economic Advisers and of the Federal Reserve Board, *inter alia*, show that the cyclical component of the higher productivity growth in the USA was negligible in recent years (see Table 19.1). These authorities ascribe 50 per cent to 70 per cent of the increase in productivity to the production and use of ICT (see Oliner and Sichel, 2000 or Baily and Lawrence, 2001). On the basis of other calculation methods, the American economist Gordon estimates, however, that nearly 40 per cent of the American productivity growth after 1995, compared to the period 1973–95, is explained by the economic cycle (see Gordon 2000). In other words, according to these calculations, about a fifth of the average annual productivity rise after 1995 is of a cyclical nature. Moreover, although by far the majority of studies are now assuming that ICT was a driving force behind the American productivity growth,

Table 19.1 Acceleration in growth of labour productivity and contributions in the USA

	Gordon (2000)	*Oliner and Sichel (2000) (Federal Reserve Board)*	*US Council of Economic Advisors (2001)*
Period considered	1995–9	1995–9	1995–2000
Reference period	1973–95	1991–5	1989–95
Acceleration in productivity growth attributable to:	1.33	1.16	1.63
Capital deepening	0.33	0.33	0.38
ICT-related	–	0.50	0.62
Other	–	–0.17	–0.23
Total factor productivity	0.31	0.80	1.19
ICT production	0.29	0.31	0.18
Rest of economy	0.02	0.49	1.00
Other factors	0.69	0.04	0.04
Cyclical effects	0.50	–	0.04
Price effects	0.14	–	–
Quality of labour	0.05	0.04	–

Sources:
Gordon, R.J. (2000) ‘Does the New Economy measure up to the great inventions of the past?’, *The Journal of Economic Perspectives* 14: 49–74.
Oliner, S. and Sichel, D. (2000) ‘The resurgence of growth in the late 1990s: Is information technology the story?’ *The Journal of Economic Perspectives* 14: 3–22.
US Council of Economic Advisers (2001) ‘Economic Report of the President’, Washington.

the channels through which the impact runs remain uncertain. The importance of ICT for the economy can be expressed in different ways, with a not always equally clear impact on the long-term productivity growth rate.

ICT can increase the growth rate of GDP and of labour productivity directly via the production of final consumer and investment goods in the ICT sectors themselves. The ICT sectors include companies which manufacture computer and telecommunications equipment or provide services relating to this equipment, such as, for example, software, mobile telephony or information provision via the Internet. The direct share of the ICT-producing sectors in the GDP amounted in 1998 to only 6.6 per cent in the USA and 4.9 per cent in the EMU, but according to calculations of the ECB in both areas they experienced the strongest growth rate of labour productivity in the 1990s. In the second half of the 1990s, the annual labour productivity growth in ICT-producing manufacturing industry in the USA and the EMU was a good ten times higher than average in the total economy. As regards the ICT-producing services, this was one and a half and six times as great respectively. In view of the relatively small scale of the ICT sectors in the economy as a whole, the effect on the overall productivity trend remained limited, however, to about one fifth in the USA and one sixth in the EMU.

The growth in labour productivity can also benefit indirectly from ICT applications by using them as a capital factor in the production process. The growth in productivity is then a consequence of a rise in the quantity of ICT capital per unit of labour deployed (i.e. the so-called 'capital deepening'). No large-scale domestic ICT sector is necessary for this, since ICT goods are also available via imports. The average annual increase in the computer-related capital investments in the USA amounted to about 18 per cent in the period 1990–5 and subsequently almost twice as much, compared to only 2 per cent and 3 per cent respectively for the other capital investments. This relative shift to ICT investments was boosted by the favourable price trend of ICT goods compared to that of other capital goods (see Figure 19.5). Since 1990, the prices of computer capital goods worldwide have declined by a factor in excess of 20, compared to a factor of only 1.5 for the prices of the other capital goods. Calculations by the ECB show, however, that the productivity dynamics of ICT investments for both the USA and the EMU have so far remained limited. The ECB approaches the importance of ICT investments through the share of the ICT-using sectors in the total economy. This share in 1998 was about three times as large in the EMU and the USA as that of the ICT-manufacturing sectors, but the growth in labour productivity there was considerably lower and after 1995 a fall was even recorded (see ECB, 2001).

Advocates of a higher long-term productivity growth point mainly to the great importance of the general dissemination of the use of new ICT. More specifically, ICT would generate positive spillover or network effects, as a result of which the overall efficiency of the entire economy (i.e. the total factor productivity) comes out higher on a lasting basis. Investments to improve and accelerate business communication (such as e-mail, e-commerce or information dissemination via the Internet) can, for example, work out favourably for several links in the production chain, because each time savings are made on transaction costs. According to some, ICT would thereby counteract economic

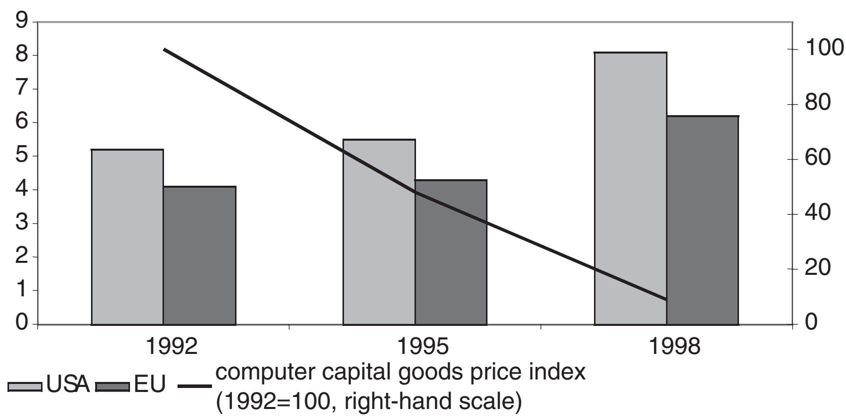


Figure 19.5 Investments in ICT (as % of GDP).
Sources: Morgan Stanley Dean Witter, OECD.

cycles, amongst other reasons because ICT-supporting logistics link up production more closely to sales, as a result of which superfluous stocks are avoided (the so-called 'just-in- time' stocks management). However, the hefty slowdown in activity in the USA since the end of 2000 has shown that ICT in reality can at the very most mitigate the cyclical volatility surrounding the growth trend. This is mainly because by definition companies cannot see economic shocks coming or at least cannot fully estimate their scale in advance. To date there is no conclusive empirical proof of the existence of substantial macroeconomic carry-over effects as a result of ICT for either the EMU or the USA. Some economists conclude that up to a third of the American productivity rise is to be attributed to sectors other than those related to ICT (see e.g. Nordhaus, 2001). On the basis of a different methodological approach, the economist Gordon mentioned above calculated, however, that spillover effects in the USA after 1995 only occurred within the consumer durables production sector, i.e. barely 12 per cent of the total American private economy. Outside this, the growth in total factor productivity even underwent a substantial decline. Calculations by the ECB too indicate that the positive carry-over effects of ICT-use in the EMU in this period were very limited or non-existent.

The fact that new ICT, compared to earlier major discoveries, such as for example the electric motor, has less clout in the total economy is probably largely attributable to a so-called 'base effect'. Since economic activity, and consequently the number of products and services, is significantly more extensive today than at the time of the earlier major discoveries, more innovations now have to be introduced per year to bring about the same growth in productivity. Another possible explanation is that ICT applications have so far yielded relatively few genuinely new products or services. For instance, in the case of Internet applications, it is quite often a matter of substitution, duplication or increasing the accessibility of already existing forms of relaxation or obtaining information. In many cases, for the supplier of services this leads simply to a sharper rise in costs than in income, or shifts take place in market shares without the total sales increasing. Nevertheless, the doubts about the existence of positive spillover effects probably also partly boil down merely to measurement errors, which obscure the picture of the productivity trend. In the past decade, the gradually increasing introduction of ICT applications mainly had effects on the services sector, where production and productivity are comparatively difficult to measure. As a result, productivity gains deriving from new technologies in part remained outside statistical observation. This was further aggravated as the economic weight of these services increased worldwide. Measurement problems also occurred because the retail price index did not always take due account of the quality of goods and services, so that price rises which were attributable to quality improvements were unjustly considered as inflation. In so far as this index overestimated inflation, the productivity rise was at the same time underestimated.

4 The role of structural characteristics

In addition to capital deepening and direct spillover effects, the use of new technologies may also bring about complementary or substitution effects on productivity indirectly. A great deal depends here on the extent to which the deployment of other factors of production changes at the same time and replacement or expansion of labour and other capital goods occurs. In practice, the use of technological progress in the production process and the way in which this influences the deployment of other factors of production is ultimately strongly determined by the structural-economic characteristics of a country. A strong indication of this is that differences in labour productivity are usually greater between countries than between sectors within countries, which indicates that explanations of the level of productivity are largely cross-sector (see e.g. Pilat, 1996 or Carree *et al.*, 1999). The characteristics alluded to relate to institutions and policy measures which determine the economic environment in which people develop skills and companies invest and produce. They are multifarious and are located *inter alia* in the field of the public infrastructure and administration (less administrative red tape, privatization of state-owned undertakings, etc.), industry (development of industrial or technology zones, reduction of subsidies to flagging industries, tax cuts, lifting of trade barriers, availability of venture capital for innovative companies, etc.) or the labour market (education and training etc.). Often they influence productivity through their impact on absolute and relative factor prices as well. Higher factor prices have a negative impact on economic activity via falling profitability of the undertakings, whilst relative price changes of the factor inputs will change the relative factor mix for a given volume of production.

In the past decade, many economists have upheld the view that macroeconomic performances are closely connected with the smooth working of the market mechanism. Thus, a *flexible labour market* would benefit the power of the economy to adjust and be accompanied by lower unemployment, but also by higher investment including that in the new technologies. This flexibility is expressed *inter alia* in flexible labour legislation, a harmonious relationship between workers and employers, low non-wage costs (such as taxes, costs with regard to recruitment and dismissals, etc.) and a strong degree of mobility of labour. Rigidities, on the other hand, render the reorganization of production processes more difficult (for example, the deployment of workers in new tasks or changing the structure of the demand for labour as a result of the need for workers with higher skills and other training) and would consequently impede firms in capitalizing rapidly on new opportunities. In conjunction with sunk investment costs, they increase the inclination to get bogged down in old technologies, even if a more efficient technology has become available in the meantime (see Saint-Gilles, 1996). Some studies compare the strong productivity gains in the USA after 1995 in association with the relatively flexible US labour market with those in Europe (see e.g. IMF, 2001). A number of generally small European countries, such as Finland, Ireland, The Netherlands and Portugal, nevertheless carried out important labour market reforms in the past decade and precisely these countries achieved productivity increases which could stand comparison with the American. Nevertheless, the positive impact of smooth working of the labour market on productivity should not be overestimated as there are also mechanisms which point more to a reverse effect. Measures which, for example, aim for more flexible wage formation make it more attractive for firms to produce labour-intensively, as a result of which the growth in labour productivity slows down. However, this labour-saving effect is mainly relevant for labour involving a low level of skills, which is more easily substituted by capital. On the other hand, there is a growing prospect of greater skills with advancing technological developments (so-called 'skill-biased technological change'), so that for this category a more flexible functioning of its labour market probably causes productivity to rise.

The relatively strong, ICT-related investment boom in the USA, compared to the EMU, was largely attributable to differences in costs of *adjustment with regard to the capital stock* of companies. In the USA, these were significantly lower, which offered American firms the advantage of applying new technological developments more quickly and on a larger scale. With high costs of adjustment, investments are adapted more slowly and are consequently less volatile. For instance, the standard deviation of the cyclically corrected level of investment (i.e. the ratio between investment and GDP) in the USA is 20 per cent higher than in Europe. However, there is a great deal of discussion and consequently uncertainty about the scale of such costs of adjustment. Estimates indicate that in Europe they account for some 13 per cent of total investment expenditures, i.e. nearly twice as much as in the USA. For investments in computer equipment, they rise to 40 per cent, which is also twice those in the USA (see Roeger, 1999). Apart from visible transaction, information or study costs, there are however also costs to be attributed to government regulations, organizational aspects within the company itself (such as, for example, training or recruitment of staff) or hindrances with regard to capital procurement through the financial markets. In reality, the costs of adjustment are therefore significantly higher, and according to the largest estimates even up to four times the expenditure on ICT investments (see Kiley, 1999). The fact that in Europe it is generally more difficult than in the USA to raise venture capital for new technologies, appears *inter alia* from the share of the ICT sectors in venture capital investments. In the USA, this share was over 50 per cent in 1998, compared to only just over 10 per cent in Europe. The number of days required to start up a new company as a consequence of administrative procedures is only one or two weeks in the USA, as against several months in some EMU countries (see also Figure 19.6).

Although a lot of competitive explanations of international productivity differences on the basis of *differences in government policy* have been put forward, there is until this moment only little conclusive empirical confirmation of it. This is largely attributable to the fact that, in addition to the existing policy situation, the changes to it, which are hard to predict, usually also play a crucial role, such as for example uncertainty about impending tax rises or new forms of government regulation. As Figure 19.7 shows, a strict employment protection legislation often goes

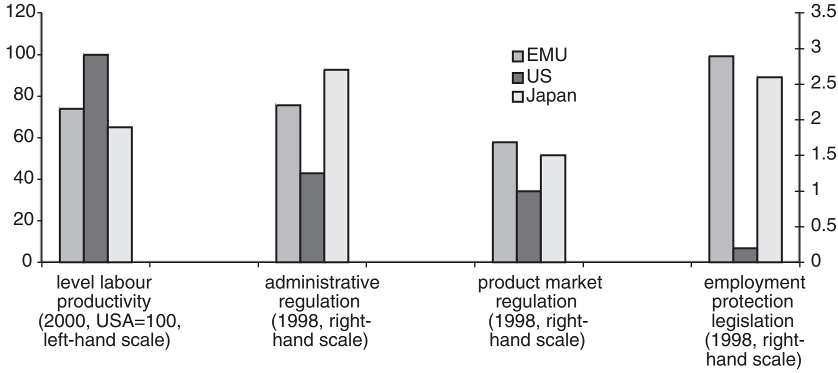


Figure 19.6 Productivity and regulation.
Sources: OECD, Nicoletti *et al.* (1999).

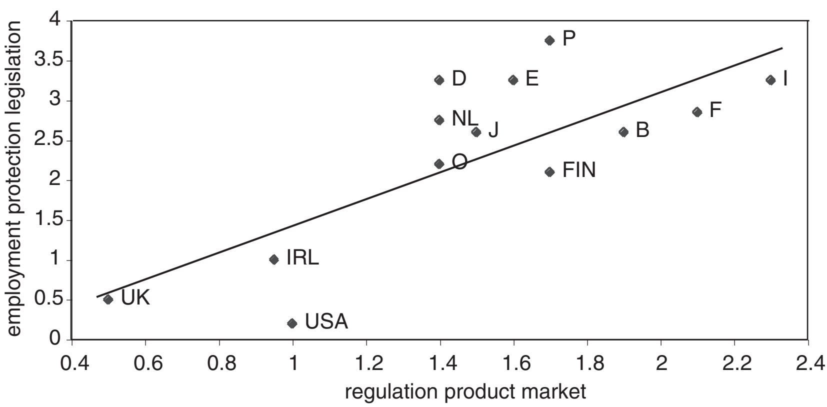


Figure 19.7 Regulation of the product and labour market (1998).
Sources: OECD, Nicoletti *et al.* (1999).

hand in hand with a rigid product market regulation, but it surely not always involves countries with a relatively unfavourable productivity level of development. Research on the subject focused mainly on investments in the public infrastructure and on the tightening up of environmental regulations. According to this, for the USA, a 1 per cent rise in the public capital stock would cause the total factor productivity to rise by 0.4 per cent. In other countries for which a causal link has been established (Germany, Canada, Belgium and Sweden), the impact admittedly appeared (a lot) smaller. In addition, the causality behind the relationship is questioned, since high growth in productivity may just as well be the driving force behind the government investments (see Aschauer, 1989; Ford and Poret, 1991; Munnell, 1993 and Everaert and Heylen, 1998). There is also lack of agreement about the harmful impact of environmental requirements on productivity. Studies for the USA indicate that 10 per cent to 30 per cent of the slowdown in productivity growth in the 1970s and 1980s was attributable to this. Probably this finding is greatly exaggerated. The environmental regulations in the USA were tightened up still further in the 1990s, which did not prevent the growth in productivity accelerating sharply then. Moreover, tightening up the environmental requirements may also boost productivity because at the same time it forces companies to rethink their systems of production (more rapidly) or because less pollution may promote other activities, such as for example tourism. Finally, traditional measures of productivity do not take account of environmental damage. Research for the USA has shown that the damage avoided through the environmental regulations was five to fifteen times greater than the costs associated with it (see Repetto *et al.,* 1997).

Intuitively, one would expect high (public and private) spending on *research and development (R&D)* to promote the production and use of new technologies and in this way benefit productivity. Nevertheless, studies so far provide no answer to the possible relationship between international productivity differences and varying efforts in the field of R&D (see e.g. Frantzen, 1998). Apart from international differences in the efficiency of domestic R&D, this can be explained by the fact that technological know-how has increasingly obtained the character of a free good in the past few decades. As a result, the industrialized countries were able to copy one another's efforts in the field of R&D or their own efforts were focused on the commercial application of existing technologies. According to OECD calculations, the use of new technologies produced abroad in the majority of industrialized countries explains half of the movement to catch up with the USA in terms of labour productivity up to 1995 (see OECD, 1996; Lichtenberg, 1992; Coe and Helpman, 1995 and Moez El Elj, 1999). Especially in small open countries, which are relatively better served by the international dissemination of new technologies, innovations devised elsewhere have contributed significantly more to the productivity increase than domestic R&D. Possibly this also explains why many of these countries, including Belgium, in spite of considerable arrears in innovation in relation to other major industrialized countries, were nevertheless able to reach a decent productivity level in recent decades (Eurostat, 1999 and Verspagen and Hollanders, 1998). For large blocs, such as the three economic superpowers, the link between productivity and R&D is more visible (see Figure 19.8).

Research has shown that *education and training* are an important explanatory factor of the level of productivity, but only play a marginal role in the long-term productivity rise (see also Figure 19.8). One year's difference in the average number of years of schooling of the population apparently gives rise to a difference in the level of labour productivity of 5 per cent to 10 per cent. The effect of an additional year of schooling on the long-term productivity increase is apparently limited to less than 0.1 per cent per year (see Englander and Gurney, 1994). On the other hand, the link between the proportion of qualified people or those who have received higher education in the total population and productivity is far less clear. The explanation for this probably lies in the fact that such figures give no insight into the quality of the education or into its tie-up with business life and consequently also because labour productivity is also influenced by on-the-job

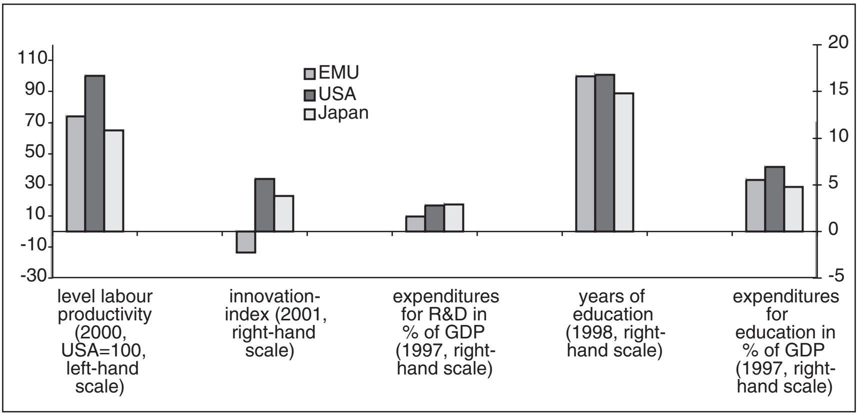


Figure 19.8 Productivity, innovation and education.
Sources: OECD, EC.

training. According to the experience of foreign establishments of multinationals, the same level of productivity can usually be obtained with local workers as that of the parent company via on-the-job training (see Baily and Gersbach, 1995). In addition, the return on investments in human capital is greater in countries which are already relatively well-provided with this, which naturally has to do with the wider possibilities to exchange ideas. Finally, it is precisely these countries which often receive an influx of foreign scientists ('brain drain').

International trade contributes to a favourable productivity trend via the dissemination of new technologies (so-called 'spillovers'), a greater use of economies of scale or the opening up of new exports markets. The internationalization of goods markets and the accompanying increased competitive pressure also require that an ever broader and better product range must be offered at the lowest cost (see e.g. Nickell, 1996, and Lever and Nieuwenhuijsen, 1998). This in turn stimulates the intensity of R&D and usually means that only the strongest firms survive. A simple calculation shows that in the OECD countries with an average annual growth in exports in the period 1980–2000 higher than the group median, the annual growth in the total factor productivity was on average 0.5 percentage point higher than in countries with a low growth in exports. For labour productivity, this was even nearly 1 percentage point higher. Hence the relatively high level of productivity in a number of small open countries (with Belgium and The Netherlands in the lead) was also probably partly attributable to the strong competitive environment which forced them to undertake cost control and product specialization (for example, the pharmaceuticals industry in Switzerland and the car industry in Belgium). Especially in the UK and Japan, a lack of competition is in turn one of the main causes of the relatively low level of productivity. Trade barriers, price controls, strict regulations (including over the use of land) and small firms in many industries limited the pressure there on firms to close up the productivity gap with the USA. The positive impact of international and domestic competition on productivity finally also emerges from studies on for example the deregulation in the telecoms sector. On the other hand, the low productivity in the German beer industry, for example, is in turn attributable to the almost exclusively local operation (see Wagner and Van Ark, 1996 and Lovegrove *et al*., 1998).

5 Policy implications

The objectives of government economic policy are varied and, in addition to achieving a satisfactory percentage of economic growth, also include the optimum utilization of the factors of production *available in the economy*. This is in contrast to firms, which aim to deploy the factors of production *used by them* as efficiently as possible. This means that a high utilization of labour (or 'full employment') for the government is not only prompted by efficiency considerations, but also by the concern to drive back poverty and social inequality. Moreover, high unemployment implies substantial budget spending for the government in the form of benefits, and lower tax and parafiscal revenue. To the extent that productivity gains therefore go hand in hand with massive redundancies, they are not desirable from the point of view of society as a whole. In practice, the boosting of labour-friendly productivity by the government is consequently no easy matter, especially as the determinants of productivity are also multifarious and often interlinked, and the causal link is sometimes unclear.

In view of these uncertainties, government policy must in the first instance be confined to the elimination of institutional and legislative obstacles so that innovative economic activities can develop to the full. The fact that the EMU countries are lagging behind the USA in terms of dynamism and technological innovation is often attributable to the still strong restriction and distortion of competition in sectors such as energy, transport and communications, to administrative red tape when starting up new businesses or to interventionism, such as the granting of special trading conditions or subsidies to traditional sectors which prevent companies rapidly switching to new high-tech production techniques or end products. In addition, the governments in the EMU must make extra efforts as regards the support of commercially oriented R&D efforts. The majority of innovations which are relevant to the growth in productivity are the result of deliberate commercially oriented R&D efforts. Especially in comparison with the USA, the European countries are still falling short as regards private investments in R&D and they apply for relatively few high-tech patents. In the nineteenth century too, some creativity on the part of the government ensured that new inventions were given a strong commercial boost and were protected (for example, the introduction of the 'patenting system' and the idea of 'limited liability'). It is generally particularly crucial to carry out R&D in the home country. The international dissemination and imitation of the results of R&D, as took root in the past decades, carry with them the risk of piracy, as a result of which the inclination of firms to engage in R&D themselves declines. Moreover, in the absence of extensive domestic R&D, a 'fished-out effect' rapidly occurs, as possibilities to imitate become exhausted. In part, these aspects of behaviour offer an explanation for the perceptible decline in the growth in investment in R&D in the OECD in the 1990s (to 1 per cent to 2 per cent on average per year, compared to 5 per cent on average in the 1980s), although the decline in public financing of R&D as a result of budgetary reform and the further increase in importance of the services sector, where the R&D spending is traditionally a bit lower, also played a role.

On the demand side of the economy too, the development of new technologies and their impact on economic growth can have a hefty impact. Modern ICT applications are 'knowledge products' *par excellence* and it is obvious that a 'learning' society accepts such innovations more rapidly than a 'conservative' one. A policy implication of this is that the government may not focus exclusively on promoting innovation in firms, but must also pay attention to strengthening the ICT knowledge and skills among the population. Traditionally, consumers only gain knowledge after purchase of a physical product which incorporates this knowledge. The government can give a helping hand here and lower the knowledge threshold by simplifying the access to new products or services, *inter alia* through education or by boosting the use of the Internet. Indirectly, the use of new technologies by the government itself in its administration and public services (e.g. PC use, Internet, etc.) can also contribute to this. In turn, an increased interest on the part of citizens will prompt companies to achieve commercially worthwhile innovations more rapidly and on a larger scale.

Appendix 19A Measurement of total factor productivity

In literature, the so-called 'Solow analysis' is generally used as a basis for the calculation of the total factor productivity. This allows the real growth of added value in an economy (*Y*) to be shared between that which is attributable to labour (*A*), to capital (*K*) and to the rest, the total factor productivity (*TFP*):

$$\frac{\Delta Y}{Y} = \alpha.\frac{\Delta A}{A} + (1-\alpha).\frac{\Delta K}{K} + \frac{\Delta TFP}{TFP}$$

where the weights α and $1-\alpha$ can be calculated as the respective shares of wages and profits in gross added value.[4] In the equation, only the term *TFP* cannot be observed. Reformulation allows the growth in total factor productivity to be described as a weighted average of the growth in labour productivity (*AP*, i.e. the production/labour ratio) and the growth in capital productivity (*KP*, i.e. the production/capital ratio). The shares in income of the factors are used as weighting coefficients:

$$\frac{\Delta TFP}{TFP} = \alpha.\frac{\Delta AP}{AP} + (1-\alpha).\frac{\Delta KP}{KP} \quad \text{and} \quad TFP = \frac{1}{(1/AP)^{\alpha}.\ (1/KP)^{1-\alpha}}$$

Changes in the total factor productivity then reflect fluctuations in the profitability of the quantities of labour and capital deployed ($\alpha.AP$ and $(1-\alpha).KP$ respectively) as a result of changes in the productivity of labour and/or capital. In view of the relatively large share of wages in added value (65 per cent to 70 per cent in the majority of industrialized countries), the total factor productivity often runs parallel to the labour productivity. The difference between the two is determined by the capital intensity of the production, i.e. the quantity of capital goods per unit of labour deployed (*KI*):

$$\frac{\Delta AP}{AP} = \frac{\Delta TFP}{TFP} + (1-\alpha).\frac{\Delta KI}{KI}$$

Considered in this way, the growth in labour productivity may be the consequence of a rise in the quantity of available capital per unit of labour or of an increase in the overall efficiency of the production process by an improvement of total factor productivity. Ideally, *TFP* should reflect the dissemination of new technologies throughout the economy. In practice, that is not always the case, since *TFP* is calculated as a residual item and therefore also reflects distortions, for example as a result of problems of measurement.

Notes

1 This was accompanied by an accommodating government policy (a hefty increase in government spending, usually low real interest rates), a stable international monetary system (Bretton-Woods system of fixed exchange rates), a professionalization of corporate policy, the expansion of multinationals and the liberalization of world trade (GATT trade rounds, establishment of the EEC).

2 In the economic literature, documented examples relate to the cooling and control system of nuclear power stations and combating plant diseases. See Cowan (1990) and Cowan and Gunby (1996).

3 The share of computers in plant equipment (excluding buildings) in the industrialized countries was estimated for the early 1990s at five per cent at the very most. See Oliner and Sichel (1994).

4 For this purpose, it is required that the factors of production are remunerated at their marginal productivity. The formula is based on a Cobb-Douglas production function with constant scale returns.

References

Aschauer, D. (1989) 'Public investment and productivity in the Group of Seven', *Economic Perspectives* 13: 17–25.

Atella, V. and Quintieri, B. (1998) 'Productivity growth and the effects of recessions', *Giornale degli Economisti e Annali di Economia* 57: 359–86.

Baily, M. and Gersbach, H. (1995) 'Efficiency in manufacturing and the need for global competition', *Brookings Papers on Economic Activity,* 307–58.

Baily, M. and Lawrence, R. (2001) 'Do we have a new e-conomy?', NBER Working Paper, 8243.

Carree, M., Klomp, L. and Thurik, A. (1999) 'Productivity convergence in OECD manufacturing industries', *Economics Letters* 66: 337–45.

Coe and Helpman (1995) 'International R&D spillovers', *European Economic Review* 39: 859–87.

Cowan, R. (1990) ‘Nuclear power reactors: a study in technological lock-in’, *Journal of Economic History*, 541–67.

Cowan, R. and Gunby, P. (1996) ‘Sprayed to death: path dependence, lock-in and pest control strategies’, *Economic Journal*, 521–42.

ECB (2001) ‘De nieuwe technologieën en de productiviteit in het eurogebied’, *Monthly Report*, July.

Englander, A. and Gurney, A. (1994) ‘La productivité dans la zone de l’OCDE: les déterminants à moyen terme’, *Revue économique de l’OCDE* 22: 63.

Eurostat (1999) *Community Innovation*, Survey 1997/1998.

Evans, C. (1992) ‘Productivity shocks and real business cycles’, *Journal of Monetary Economics* 29: 191–208.

Everaert, G. and Heylen, F. (1998) ‘Public capital and productivity growth in Belgium’, Working Paper, University of Ghent.

Ford, R. and Poret, P. (1991) ‘Infrastructure and private sector productivity’, *OECD Economic Studies,* 17: 63–89.

Frantzen, D. (1998) ‘Onderzoek en ontwikkeling, buitenlandse technologische diffusie en productiviteitsgroei: een internationale cross-sectieanalyse’, *Economisch en Sociaal Tijdschrift*, 69–96.

Gordon, R. (2000) ‘Does the New Economy measure up to the great inventions of the past’, *The Journal of Economic Perspectives* 14: 49–74.

IMF (2001) *World Economic Outlook*, October (Chapter III: ‘The information technology revolution’, 103–42).

Kiley, M. (1999) ‘Computers and growth with costs of adjustment: will the future look like the past?’, Board of Governors of the Federal Reserve System Finance and Economics Discussion Series, 36.

Lever, M. and Nieuwenhuijsen, H. (1998) ‘Concurrentie verhoogt productiviteit’, *ESB*, 744–6.

Lichtenberg, F. (1992) ‘R&D investment and international productivity differences’, NBER Working Paper, 4161.

Lovegrove, N. *et al.* (1998) ‘Why is labor productivity in the United Kingdom so low?’, *The McKinsey Quarterly*, 4: 45–57.

Moez El Elj (1999) ‘Trade, transmission of technology and productivity: an empirical analysis for OECD countries’, *Revue d'économie politique* 2: 221–40.

Munnell, A. (1993) *Les Investissements d'infrastructure: Évaluation de leurs Tendances Actuelles et de leurs Effets Économiques*. Paris: OECD.

Nickell, S. (1996) ‘Competition and corporate performance’, *Journal of Political Economy* 4: 724–47.

Nicoletti, G., Scarpetta, S. and Boyland, O. (1999) ‘Summary indicators of product market regulation with an extension to employment protection legislation’, OECD Working Paper no. 226, Paris.

Nordhaus, W. (2001) ‘Productivity growth and the New Economy’, NBER Working Paper, 8096.

OECD (1991) *Technology and Productivity: The Challenge for Economic Policy*. Paris: OECD.

OECD (1996) *Technology and Industrial Performance: Technology Diffusion, Productivity, Employment and Skills, International Competitiveness*. Paris: OECD.

Oliner, S. and Sichel, D. (1994) 'Computers and output growth revisited: how big is the puzzle?', *Brookings Papers on Economic Activity* 2: 273–334.

Oliner, S. and Sichel, D. (2000) 'The resurgence of growth in the late 1990s: is information technology the story?', *The Journal of Economic Perspectives* 14: 3–22.

Pilat, G. (1996) 'Concurrence, productivité et efficience', *Revue Economique* 27: 124–33.

Repetto, R., Rothman, D., Faeth, P. and Austin, D. (1997) 'Has environmental protection really reduced productivity growth?', *Challenge* 40.

Roeger, W. (1999) 'How important are nominal rigidities for understanding business cycles in Germany and the US?', *Zeitschrift für Konjunkturpolitik*, 138–51.

Saint-Gilles, G. (1996) 'Employment protection, international specialization and innovation', CEPR Discussion Paper, 1338.

Verspagen, B. and Hollanders, H. (1998) 'De Nederlandse innovatie-achterstand', *ESB*.

Wagner, K. and Van Ark, B. (eds) (1996) *International Productivity Differences, Measurement and Explanations*. Amsterdam: North-Holland.

Index