Issues in Positive Political Economy

Edited by S. Mansoob Murshed

London and New York

Issues in Positive Political Economy

Political economy – the original name for economics in its entirety – has in recent years witnessed a semantic broadening to include some of the preoccupations of classical economics. This intriguing collection of contributed work is concerned mainly with developments in the neoclassical tradition and examines the role played by rational choice in the decision-making processes of firms and the State.

Among the topics covered in this timely book are:

- the problems of the undemocratic centralisation of power in European politics
- the causes of contemporary conflict and civil war
- the problems facing economics in transition, for example Russia
- the problems of electoral uncertainty and why governments can be fiscally irresponsible

With contributions from leading scholars including Tony Addison, Richard Auty, Bruno Frey, Mansoob Murshed and Alan Rugman, this book is an important addition to the field of Political Economy and should become essential reading for all economists as well as policy makers.

S. Mansoob Murshed is at the Institute of Social Studies, the Hague, in the Netherlands.

Routledge Frontiers of Political Economy

1 **Equilibrium Versus Understanding**
Towards the rehumanization of economics within social theory
Mark Addleson

2 **Evolution, Order and Complexity**
Edited by Elias L. Khalil and Kenneth E. Boulding

3 **Interactions in Political Economy**
Malvern after ten years
Edited by Steven Pressman

4 **The End of Economics**
Michael Perelman

5 **Probability in Economics**
Omar F. Hamouda and Robin Rowley

6 **Capital Controversy, Post Keynesian Economics and the History of Economics**
Essays in honour of Geoff Harcourt, volume one
Edited by Philip Arestis, Gabriel Palma and Malcolm Sawyer

7 **Markets, Unemployment and Economic Policy**
Essays in honour of Geoff Harcourt, volume two
Edited by Philip Arestis, Gabriel Palma and Malcolm Sawyer

8 **Social Economy**
The logic of capitalist development
Clark Everling

9 **New Keynesian Economics/Post Keynesian Alternatives**
Edited by Roy J. Rotheim

10 **The Representative Agent in Macroeconomics**
James E. Hartley

11 **Borderlands of Economics**
Essays in honour of Daniel R. Fusfeld
Edited by Nahid Aslanbeigui and Young Back Choi

12 **Value, Distribution and Capital**
Essays in honour of Pierangelo Garegnani
Edited by Gary Mongiovi and Fabio Petri

13 **The Economics of Science**
Methodology and epistemology as if economics really mattered
James R. Wible

14 **Competitiveness, Localised Learning and Regional Development**
Specialisation and prosperity in small open economies
Peter Maskell, Heikki Eskelinen, Ingjaldur Hannibalsson, Anders Malmberg and Eirik Vatne

15 **Labour Market Theory**
A constructive reassessment
Ben J. Fine

16 **Women and European Employment**
Jill Rubery, Mark Smith, Colette Fagan and Damian Grimshaw

17 **Explorations in Economic Methodology**
From Lakatos to empirical philosophy of science
Roger Backhouse

18 **Subjectivity in Political Economy**
Essays on wanting and choosing
David P. Levine

19 **The Political Economy of Middle East Peace**
The impact of competing trade agendas
Edited by J.W. Wright, Jnr

20 **The Active Consumer**
Novelty and surprise in consumer choice
Edited by Marina Bianchi

21 **Subjectivism and Economic Analysis**
Essays in memory of Ludwig Lachmann
Edited by Roger Koppl and Gary Mongiovi

22 **Themes in Post-Keynesian Economics**
Essays in honour of Geoff Harcourt, volume three
Edited by Peter Kriesler and Claudio Sardoni

23 **The Dynamics of Technological Knowledge**
Cristiano Antonelli

24 **The Political Economy of Diet, Health and Food Policy**
Ben J. Fine

25 **The End of Finance**
Capital market inflation, financial derivatives and pension fund capitalism
Jan Toporowski

26 **Political Economy and the New Capitalism**
Edited by Jan Toporowski

27 **Growth Theory**
A philosophical perspective
Patricia Northover

28 **The Political Economy of the Small Firm**
Edited by Charlie Dannreuther

29 Hahn and Economic Methodology
Edited by Thomas Boylan and Paschal F. O'Gorman

30 Gender, Growth and Trade
The miracle economies of the postwar years
David Kucera

31 Normative Political Economy
Subjective freedom, the market and the state
David Levine

32 Economist with a Public Purpose
Essays in honour of John Kenneth Galbraith
Edited by Michael Keaney

33 The Theory of Unemployment
Michel De Vroey

34 The Fundamental Institutions of Capitalism
Ernesto Screpanti

35 Transcending Transaction
The search for self-generating markets
Alan Shipman

36 Power in Business and the State
An historical analysis of its concentration
Frank Bealey

37 Editing Economics
Essays in honour of Mark Perlman
Hank Lim, Ungsuh K. Park and Geoff Harcourt

38 Money, Macroeconomics and Keynes
Essays in honour of Victoria Chick, volume one
Philip Arestis, Meghnad Desai and Sheila Dow

39 Methodology, Microeconomics and Keynes
Essays in honour of Victoria Chick, volume two
Philip Arestis, Meghnad Desai and Sheila Dow

40 Market Drive and Governance
Reexamining the rules for economic and commercial contest
Ralf Boscheck

41 The Value of Marx
Political economy for contemporary capitalism
Alfredo Saad-Filho

42 Issues in Positive Political Economy
Edited by S. Mansoob Murshed

Issues in Positive Political Economy

Edited by S. Mansoob Murshed

London and New York

First published 2002
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
A catalog record for this book has been requested

ISBN 0-203-16730-9 Master e-book ISBN

ISBN 0-203-26213-1 (Adobe eReader Format)
ISBN 0–415–25990–8 (Print Edition)

Dedicated to Mary Conaghan
From Mansoob

Contents

Illustrations

Tables

Figure

Contributors

Tony Addison, World Institute for Development Economics Research (WIDER), Helsinki, Finland.

Harvey W. Armstrong, Department of Geography, University of Sheffield, UK

Richard M. Auty, Department of Geography, University of Lancaster, UK.

Jitendralal Borkakoti, Middlesex University, Queensway, Enfield, UK.

Jagjit S. Chadha, Clare College, Cambridge, UK.

Paola Conconi, Department of Economics, and Centre for the Study of Globalization and Regionalism, Coventry, UK.

George Economides, University of Cyprus, and Athens University of Economics and Business, Greece.

Bruno S. Frey, University of Zurich, Institute for Empirical Economics, Switzerland.

Holger Görg, School of Economics, University of Nottingham, UK.

John Kirton, Munk Centre for International Studies, University of Toronto, Canada

Natasha Miaouli, Athens University of Economics and Business, Greece.

S. Mansoob Murshed, Institute for Social Studies (ISS), The Hague, Netherlands.

Muriel Niederle, Department of Economics, Harvard University, Cambridge, MA, USA.

Apostolis Philippopoulos, Athens University of Economics and Business, and CESifo, Greece.

Robert Read, Department of Economics, University of Lancaster, UK.

Alan M. Rugman, Kelley School of Busines, Indiana University and Templeton College, Oxford University, UK.

Martin Summer, Economic Studies Division, Österreichische Nationalbank, A-1011 Vienna, Austria.

Abbreviations

AOSIS	Association of Small Island States
APEC	Asia Pacific Economic Cooperation
BSD	Bahamas dollar
BZD	Belize dollar
CAR	Central African Republic
CARICOM	Caribbean Community
CE	competitive equilibrium
CEC	Commission for Environmental Cooperation (of NAFTA)
CEE	Central and Eastern Europe and the Baltic States
CFA	Communauté Financière Africaine
CHF	Swiss franc
CIS	Commonwealth of Independent States
CRC	Costa Rican dollar
CYP	Cypriot pound
DRC	Democratic Republic of Congo
EBRD	European Bank for Reconstruction and Development
ENGO	Environmental Non-Governmental Organization
ESP	Spanish peseta
EU	European Union
FDI	foreign direct investment
FJD	Fijian dollar
FOCJ	functional overlapping competing jurisdictions
FRF	French franc
FTA	free trade area
FTAA	Free Trade Area of the Americas
GATS	General Agreement on Trade in Services
GATT	General Agreement on Tariffs and Trade
GBP	British pound
GDP	gross domestic product
GMD	Gambian dalasi
GNP	gross national product
HCI	heavy and chemical industry
HOS	Hecksher–Ohlin–Samuelson
IESG	International Economics Study Group
IIT	intra-industry trade
IMF	International Monetary Fund
IPT	inward processing trade
IRS	internal revenue service

ISK	Icelandic krona
LDC	less developed country
LSL	Lesotho loti
MFN	most favoured nation
MIRAB	migration, remittances and barter
MMT	methylcyclopentadienyl manganese tricarbonyl
MNC	multinational company
MNE	multinational enterprises
MPLA	Movimento Popular de Libertação de Angola (Popular Movement for the Liberation of Angola)
NAAEC	North American Agreement on Environmental Cooperation
NAFTA	North American Free Trade Agreement
NEER	nominal effective exchange rates
NIE	newly industrialising economies
NGO	non-governmental organisations
OECD	Organisation for Economic Cooperation and Development
OPT	outward processing trade
PAB	Panamanian balboa
PPP	purchasing power parity
PGE	political general equilibrium
R&D	research and development
REER	real effective exchange rates
SADEC	South African Development Community
SDR	Special Drawing Right
SIDS	small island developing states
SITC	standard international trade classification
SOE	state-owned enterprise
TRIM	Trade-Related Investment Measures
TRIP	Trade-Related Intellectual Property Rights
TTD	Trinidad and Tobago dollar
UHT	ultra high temperature
UK	United Kingdom
UN	United Nations
UNCTAD	United Nations Conference on Trade and Development
UNITA	União Nacional para a Independência Total de Angola (National Union for the Total Independence of Angola)
UNU/WIDER	United Nations University/World Institute for Development Economics Research
US	United States
USD	United States dollar
WST	Western Samoa tala
WTO	World Trade Organisation
XAF	Communauté Financière Africaine franc
XCD	East Caribbean dollar
ZAR	South African rand

1 Introduction

S. Mansoob Murshed

The term political economy has several meanings (Schumpeter, 1954). In classical economics, the term was employed to cover various aspects of economic theory as well as economic policy (Mill, 1848). With the advent of the marginalist revolution in neoclassical economic analysis, political economy gradually became more concerned with broader issues and power relations. Economic theory became more 'positive' in nature, utilising mathematical methods to analyse issues in the context of optimisation problems and models of the economy or its constituent sectors. The normative side of economics was concerned with the best means of achieving an economic or social policy goal, such as the redistribution of income or the financing of a welfare programme, but the aim itself was exogenously given from the 'outside' through some unspecified political process.

The radical and Marxist tradition in economics continued utilising the term political economy, see for example Dobb (1937), reflecting the importance they attached to power relations between classes and states. Another area, where the expression continued in vogue was in the nascent subject of development economics, particularly in facets of that field which dealt with the relations between the more powerful, industrial, ex-colonial powers and the less-developed, primary producing ex-colonies. A well-known exemplar of this tradition is the Latin American focused dependency theory, see Furtado (1972) and Frank (1975), for example. Other works in the area that are more centred on South Asia, include Bagchi (1982). A recent, more historically based study, on the creation of the present developing or third world in the late nineteenth century can be found in Davis (2001). The left, however, did not enjoy exclusive copyright over the use of the term political economy; the radical right represented by the libertarian school (such as Buchanan, 1988) also used the expression in matters such as the design of the economic constitution.

Meanwhile, there have been developments in the more technical and theoretical side of neoclassical economics in the last thirty years. These have bought economic theory and political economy closer to one another. This has occurred because of two reasons. First, economic theory is increasingly seeing policy objectives as being determined within an economic model, and not given from the outside.

Second, the discourse and methodology of theoretical economics is spreading to other disciplines, especially political science. Quantification and the use of economic data is nowadays also widespread in the other social sciences.

Moving on to the recent innovations in economic theory, there is now a greater emphasis on analysing strategic behaviour (say between two rival firms), as opposed to the more traditional atomistic behaviour (as in the paradigm of perfect competition). The methodology employed involves game theory.

Another development has been the advent of the theory of asymmetric information, which attracted the 2001 Nobel Prize in economics, exemplified by the work of Akerlof (1970), Spence (1973) and Rothschild and Stiglitz (1976). The emphasis is on information that is private, and not publicly known, leading to problems of moral hazard and adverse selection. Moral hazard implies the existence of an unverifiable input of effort, necessary to carrying out a task. Adverse selection refers to information about the type of the agent, information that is private to the individual concerned. Moral hazard and adverse selection lead to difficulties in managing uncertainty, market failure, as well as problems in designing incentive-based contracts. Both moral hazard and adverse selection are said to arise in principal–agent relationships and in contract design, where the agent undertakes a task on behalf of the principal. Collectively, therefore, they are known as the problems of agency. Examples include insurance markets, risk sharing between an owner and a franchisee, and in the general design of incentive contracts based on verifiable outcomes or performance in carrying out tasks.

These problems not only exist in the private sector, but in government as well (see Laffont and Tirole, 1993). Many areas of economic policymaking, as well as a variety of bureaucratic functions including regulation, supervision and ensuring probity become problematic because of moral hazard and adverse selection. Moreover, the presence of a multiplicity of tasks and having the need to answer to several masters (the legislature, pressure groups and so on) further compounds the difficulties associated with bureaucratic functioning (see Dixit, 1996, 1999). Also, the incentives that motivate in government, academia and elsewhere, are not always pecuniary (extrinsic), but also include non-pecuniary (intrinsic) rewards such as future career prospects and the regard of colleagues (Tirole, 1994).

Furthermore, neoclassical economics has also begun to grapple with problems of sustaining commitment to optimal policies, such as inflation control, in the face of temptations to renege on prior commitments. It is not enough, for example, to have a commitment from government to a policy of low inflation. Such an announcement must be credible, and the government must be perceived to honour its commitments. Similarly, it is not sufficient to grant concessions to attract capital inflows and investment, if the outside world views these policies as temporary posturing.

In modern economic theory, governments are no longer regarded as the benevolent state which maximises social welfare, but are increasingly seen as made up of a variety of self-seeking and partisan interests. One example is the theory of lobbying and political contributions. The rigorous analysis of rent-seeking lobbies, who attempt to influence government policy through political contributions, can be found in Grossman and Helpman (1996).

In development economics, the experience of the economies in transition in Central and Eastern Europe, led to the explicit analysis of the allocation of entrepreneurial talent between production and corruption, in Murphy, Shleifer and Vishny (1991). Societies where individual incentives are tilted towards rent-seeking activities rather than productive functions will ultimately experience a growth collapse (Murphy, Shleifer and Vishny, 1993). This is all the more likely in poorly-governed societies, where the institutions of contract enforcement and the rule of law are degraded. More fundamentally, there is the realisation that devising incentive contracts is not simply a matter of covering the usual eventualities, but mechanisms for contract enforcement, such as a functioning legal system, have to be in place and cannot always be taken for granted.

All of these developments taken together imply that most economic theorists no longer regard economic policymaking as exogenous, but the endogenous outcome of a political process. It is important to understand the rules of the game determining these processes. Political processes involve opportunistic behaviour; therefore present-day commitments are not always sustainable in the future. Opportunism entails costs. This has led to the development of the notion of transaction-cost politics (Dixit, 1996, 1999). Furthermore, interventions in the policymaking process can be shallow or deep (Dixit, 1999). The former implies tinkering with the system; the latter implies altering the process and rules of the game, constitutional reform and mechanism design.

In development economics, the catastrophic income compressions in the erstwhile communist countries, combined with huge increases in inequality and poverty there, as well as the negative growth performance in Africa, and the prevalence of civil wars in these regions has led most economists to emphasise the role of institutions (pioneered in North, 1990), such as well functioning institutions of dispute settlement, law enforcement, voice, accountability and bureaucratic quality (see Campos and Nugent, 1999). The importance of institutions is also highlighted by relative success stories, such as in East Asia.

The chapters in this books will be concerned with these recent developments, mainly in the neoclassical tradition, which we may term as positive political economy. Positive because the underlying processes, explicitly or implicitly, involves rational choice in a setting of strategic behaviour. Political economy because the concern is with policies, institutions and change. The chapters represent a selection of presentations at the UK International Economics Study Group (IESG)[1] Conference in Political Economy at the University of Warwick, 15–16 April 1999, along with a few related chpaters that have, in one form or another, been presentedas papers at various IESG seminars.

Chapters 2 to 5 address issues of systemic failure. Chapter 2 by Bruno Frey proposes a new federalist structure for Europe based on local demands and democracy. In a sense, it proposes a change to the economic constitution akin to the ideas of Buchanan (1988, for example). The chapter takes the emergence of federal structures at the central level of the European Commission as its point of reference, and addresses the problems of the undemocratic centralisation of power at that level. It proposes a decentralised system of public goods provision based on

functional overlapping competing jurisdictions (FOCJ). Any group of individuals, or community, might belong to separate FOCJ for different purposes such as health, education, supply of utilities and so on. It is argued that the greatest advantage of such a system will be its competitive character, as states or existing jurisdictions no longer enjoy monopoly powers of service provision. Additionally, it would be democratic, cost efficient and truly focus on individual needs.

Chapter 3 by Addison and Murshed is on the causes of contemporary civil wars. Civil war, the most common form of contemporary warfare, has a devastating effect on economic development. Therefore, development policy cannot afford to ignore the conflict dimension. The authors argue that the relative deprivation of clearly identifiable groups in the form of unequal access to public spending, government jobs and the distribution of assets play a major part in creating the initial conditions for conflict. They also point out that ethnicity, based on religion, race or tribal grouping, can be a powerful organising principle for revolt and collective action, far stronger than other forms of solidarity such as class. Although the desire to control capturable natural resource rents, such as alluvial diamonds, by warlords may explain many contemporary civil wars, underlying grievances also play a major part. Grievances may reflect historical injustices against well-defined groups. But it could also be related to the failure of the state to provide security and public services, or the state's repudiation of an earlier contract to redistribute and treat all communities fairly. Excluded groups are forced to rely on their own ethnic ties for security and minimal social provision. There is then a situation where open conflict could erupt, given the appropriate catalyst. Poverty too fuels civil war by lowering the 'opportunity cost' of being a soldier and participating in warfare.

The chapter also attempts to explain why commitments to peace are not credible, and parties often renege on peace agreements. This can occur if one side to the conflict actually benefits from the spoils of war, diamonds, say, and is impatient to control these. Reconstruction strategies not only require broad-based development to help rebuild the social contract, but reducing intra-group inequalities may be equally important.

Chapter 4 by Auty is about Russia. It is a 'classic' mineral economy, along the lines of point-source natural resource endowment discussed in Chapter 3 by Addison and Murshed. Point-source refers to a type of natural resource that is concentrated in ownership and production, mainly in minerals. This type of endowment is in contrast to diffuse natural resources, such as with agricultural commodities. The chapter cites models of mineral-driven economic development to explain Russia's disappointing transition to a market economy. These models predict that resource abundance renders the transition more difficult, basically because the natural resource rents ease pressure for reform. More specifically, mineral rents are likely to retard both institutional and economic reform, slow economic restructuring, feed rent-seeking behaviour and corruption, and delay the resumption of sustained rapid economic growth. The analysis of Murphy, Shleifer and Vishny (1991, 1993) is applicable here in connection with corruption and the misallocation of talent in the direction of rent seeking. Certainly, the Russian government used hydrocarbon rents to co-opt opponents of reform and in so doing it institutionalised soft budgets.

This stalled tax reform and unintentionally strengthened the real exchange rate so that structural change and economic growth were both retarded. The 1998 financial crisis coincided with a negative oil shock that depreciated the real exchange rate and eased the growth constraints. However, Auty argues that the respite is likely to be temporary, without the deep institutional reforms alluded to earlier.

Chapter 5 by Economides, Miaouli and Philippopoulos formalises the link between electoral uncertainty, fiscal policy and economic growth. It is very much in the tradition of the trade-off between market failure and policy failure. The former requires government intervention, the latter concerns faulty government intervention. They utilise a formal general equilibrium model of optimal growth and fiscal policy, in which fiscal policy is endogenised through a game between two political parties that can alternate in power. The elected party uses income taxes to finance the provision of public consumption services. The model in the chapter solves for a Markov-perfect general equilibrium in Nash strategies between the political parties, and explains how electoral uncertainty, in the form of ex-ante re-election probabilities, affects the conduct of fiscal policy, and in turn how fiscal policy affects private investment and economic growth. Low re-election probabilities can induce policymakers to follow shortsighted, inefficient policies; the inefficiency here takes the form of a relatively large public sector with short-term benefits that retard economic growth. The solutions to the problem of inefficient government behaviour requires changing the rules of the game, deep systemic changes, or in the language of game theory, 'mechanism design'.

Armstrong and Read (Chapter 6) are concerned with the political economy of growth and trade strategies in small states. The second half of the twentieth century saw a substantial rise in the number of small states in the global economy, as a result of both decolonisation and the disintegration of larger states. The strong growth performance of many of these small states suggests that the analysis of the determinants of the economic performance of small states and improved understanding of their growth strategies may generate useful insights to other, larger, developing states. The more successful small states have shown themselves to be adept at identifying and taking advantage of opportunities to engage in free-riding and rent-seeking activities in the global economy, made possible by virtue of their small size.

In the same vein, Chadha (Chapter 7), analyses the choice of exchange rate regime (fixed or flexible) by small states. Flexible exchange rates are preferable in the presence of stickiness in domestic wages and prices, as they can insulate the economy from external real shocks. On the other hand, fixed exchange rates increase the credibility of anti-inflation policies by signalling commitment to low inflation. Current practice suggests that flexibility in exchange rates is favoured by most developing countries. But this does not appear to be the case for small states; the majority of which maintain some formal link to an external currency. But as these countries have few rigidities; prices seem as flexible as nominal exchange rates, there is no great loss, in output stabilisation terms from the choice of regime. Fixed exchange rates, therefore, are most likely retained because of inertia, or for some political or strategic gain.

Chapters 8 to 11 are concerned with the pattern of international trade and trade policy by firms and governments. Intra-industry and intra-firm trade is analysed by Görg in Chapter 8. In particular, the chapter is concerned with fragmentation. This refers to splitting up the production process into various components following technological progress. Plants can outsource intermediate inputs, which enhances trade in components and encourages the emergence of multinational companies that trade intermediate goods between different plants within the same firm (inward processing trade, IPT). Görg analyses fragmentation from the viewpoint of transaction costs theory. Countries attract IPT in those sectors in which they have a comparative advantage. Thus, countries which have a comparative advantage in low (high) skill intensive production will specialise in the production of low (high) skill intensive fragments. For countries at the periphery of the EU, the empirical results in the chapter show that the presence of US foreign direct investment encourages US IPT. This suggests that at least part of the inward processing trade going into these countries is intra-firm trade between different plants of the same company rather than trade between different firms. Fragmentation may not help developing countries to build up a developed competitive industrial structure, unless it is also accompanied by foreign direct investment (FDI) inflows.

Chapter 9 by Borkakoti is also concerned with technical progress. He argues that technological progress is clearly the prime force behind the phenomenal expansion of world trade during the last two centuries. Theories, which put forward factor endowments (Heckscher–Ohlin) or labour productivity (Ricardian) as the crucial determinant of international trade, have technological innovation implicitly functioning in the models. The new trade theories, however, redress the balance by emphasising product innovation and imitation as driving international trade, particularly intra-industry trade.

Chapter 10 by Rugman and Kirton, analyses the North American Free Trade Agreement (NAFTA) from the point of view of corporate strategies of multinational enterprises as they respond to new trans-border environmental regulations. NAFTA is the first international trade and investment agreement to incorporate environmental laws and standards. As the new environmental and trade regime of NAFTA emerges the authors find that the potential set of relevant strategies for firms has expanded, and a situation of complex institutional responsiveness has developed.

Conconi in Chapter 11 is also concerned with environmental regulations and trade policy. But the methodology she employs is different. She uses a common agency model, where many principals deal with the same agent. The model also analyses the formation of green and producer lobbies who make political contributions to influence government policy. Will these lobbies be allies or rivals in their attempts to sway the government? The outcomes depend on whether or not governments have both trade (import tax) and environmental (pollution tax) policies at their disposal, and a taxonomy of results are presented in the chapter.

Chapter 12 by Summer and Niederle deal with problems of income tax auditing and enforcement. These difficulties arise from the fact that auditing of taxpayers has to be done by employees of the enforcement authority who may not be motivated to

act in its best interest. A form of auditor moral hazard, where the auditor can shirk on his auditing effort and claim to have found no evidence of irregular behaviour of taxpayers, is analysed in the model. This would make him ineffective as an enforcement instrument. A simple incentive scheme that motivates the auditor to properly investigate income tax files is studied. The authors demonstrate that such a scheme can lead to a complex interaction between auditor's and taxpayer's incentives making it difficult for the authority to achieve effective enforcement. They argue that the economics of tax enforcement might have to consider other forms of intrinsic or non-monetary motivation.

Note

1 This group receives financial support from the UK Economic and Social Research Council (ESRC).

References

Akerlof, G. (1970) 'The market for lemons: qualitative uncertainty and the market mechanism', *Quarterly Journal of Economics*, 84, 488–500.

Bagchi, A. K. (1982) *The Political Economy of Underdevelopment*, Cambridge: Cambridge University Press.

Buchanan, J. (1988) 'Contractarian political economy and constitutional interpretation', *American Economic Review*, 78, 135–9.

Campos, N. F. and J. B. Nugent (1999) 'Development performance and the institutions of governance: evidence from East Asia and Latin America', *World Development*, 27, 439–52.

Davis, M. (2001) *Late Victorian Holocausts: El Niño Famines and the Making of the Third World*, London: Verso.

Dixit, A.K. (1996) *The Making of Economic Policy: A Transaction-Cost Politics Perspective*, London: MIT Press.

Dixit, A.K. (1999), 'Some lessons from transaction-cost politics for less-developed countries', unpublished manuscript, Princeton University.

Dobb, M.H. (1937) *Political Economy and Capitalism*, London: G. Routledge & Sons.

Frank, A.G. (1975) *Capitalism and Underdevelopment in Latin America*, Harmondsworth: Penguin.

Furtado, C. (1972) *The Economic Development of Latin America*, Cambridge: Cambridge University Press.

Grossman, G.M. and E. Helpman (1996) 'Electoral competition and special interest politics', *Review of Economic Studies*, 63, 265–86.

Laffont, J-J and J. Tirole (1993) *A Theory of Incentives in Procurement and Regulation*, London: MIT Press.

Mill, J.S. (1848) *Principles of Political Economy* (reprinted 1994), Oxford: Oxford University Press.

Murphy, K., A. Shleifer and R. Vishny (1991) 'The allocation of talent: implications for growth', *Quarterly Journal of Economics*, 106, 503–30.

Murphy, K., A. Shleifer and R. Vishny (1993) 'Why is rent seeking so costly to growth', *American Economic Review*, 83, 409–14.

North, D.C. (1990) *Institutions, Institutional Change and Economic Performance*, Cambridge: Cambridge University Press.

Rothschild, M. and J. Stiglitz (1976) 'Equilibrium in competitive insurance markets: an essay on the economics of imperfect information', *Quarterly Journal of Economics*, 90, 629–50.

Schumpeter, J.A. (1954) *History of Economic Analysis*, London: George Allen and Unwin.
Spence, M. (1973) 'Job market signalling', *Quarterly Journal of Economics*, 87, 355–74.
Tirole, J. (1994) 'The internal organization of government', *Oxford Economic Papers*, 46, 1–29.

2 A proposal for a new dynamic federalism in Europe

Bruno S. Frey

The coming together of Europe is a fascinating and wonderful development. It seeks to overcome the perennial strife between the European nations whose worst manifestations were the two world wars of the last century. It should also solve the ongoing civil wars of long duration and bitterness that still plague this continent: in the former Yugoslavia, Ireland, the Basque country and Corsica. European history is not just a catalogue of belligerent behaviour. Europe can be proud of its achievements in the arts, sciences, and in its way of life. The basis of it all is *diversity*: a Scot is unlike a Sicilian, a Breton unlike a Bavarian, and an Andalusian unlike a Åland islander, and so on.

The vision developed in this chapter is based on these two basic ingredients: the future Europe has to be peaceful, and maintain its diversity. The proposal of democratic decentralised jurisdictions drastically differs from the European Union existing today. In particular, it emphasises the citizen in the political process, and proposes to decentralise the political process to the functionally most appropriate level. The proposal thus seeks to redress the two well-known and often lamented shortcomings of the European Union: its 'democractic deficit' and its 'decentralisation deficit'. The proposal in this chapter also differs markedly from other reform plans, seeking to mitigate the two deficits, which are currently being discussed.

The favoured answer to the democratic deficit is said to lie in further strengthening the European Parliament. But this measure does not necessarily engage the citizens of the European Union (EU) more fully in the political process. Rather, it may even increase the distance between the citizens and the decision-making body (Strasbourg or Brussels are even more remote than the national capital). Moreover, the concentration of over 600 professional politicians tends to invite the formation of yet another '*classe politique*', now at the European level, similar to the 'Eurocrat' counterpart of national bureaucrats.

The current plan to mitigate the 'decentralisation deficit' does not fare any better. It invokes the principal of subsidiarity, something that is ineffective as long as the regions of Europe financially depend on Brussels and the central governments of their nations. Political decentralisation requires the power to tax at lower echelons of government. And only then are they induced and able to balance the

benefits and costs of public expenditure. Only then is there fiscal responsibility, and only then will politicians be motivated to use scarce resources for the benefit of the citizenry at the grass-root level.

The vision proposed here is quite radical, and perhaps difficult to accept, but not outlandish:

1. The proposal of democratic decentralised jurisdictions is based on concepts central to economics, and in particular to the economic theory of federalism (Bird 1993; Breton 1996), for example, 'fiscal equivalence', 'voting by foot' or 'clubs'. However, they are combined in new ways, yielding a different type of federalism.

2. The proposal can be put into reality. Indeed, there are pertinent examples in history. Most importantly, the proposal can be introduced gradually.

3. The proposal does not require the dismantling of the nation states forming the European Union. Though nationalism has done enormous harm to Europe in the twentieth century, it is still a powerful force, and there is little sense in directly attacking it. But what is proposed is that there should be other jurisdictions besides it, and that the nation's right of existence has to be demonstrated by its effectiveness to efficiently care for the preferences of the population.

The proposal should not be understood to be a wholesale critique of the European Union. Rather, European integration has been very successful in opening markets within its confines. The four freedoms of liberal trade with respect to goods, services, capital and labour have been achieved to a considerable degree. Though protectionist tendencies still exist and make themselves felt almost daily, the European Union can be proud of having achieved a free market covering almost the whole of Western Europe.

The European Union is, however, not just a success story, but also one of failure. A flawed concept of Europe has increasingly taken over, and the unification process has taken a wrong turn. While these tendencies have accumulated over time, they have become dominant recently. This mistaken concept of Europe consists in identifying integration with *homogenisation* and *harmonisation*. There are hundreds of laws and directives in the European Union working in this direction. But the essence of Europe is its diversity. The strength of Europe is its wide variety of ideas, cultures and policies. Diversity, and not unity, has been the crucial element of Europe's rise in history and continues to be so. A homogenised Europe loses its *raison d'être*, and will lose its economic and political role.

Integration should serve to foster this variability. It should set the rules under which the strength of the manifold components of Europe can develop. Opening up economic markets for free trade is exactly such a beneficial rule: it allows suppliers to specialise in the production of differentiated goods and services following the law of comparative advantage. However, no such open and competitive market for politics has been established. On the contrary: competition between

governments was successfully restricted by the various European treaties and institutions. No steps have been undertaken to actively institutionalise competition between governmental units at all levels. Welfare can be improved substantially by promoting competition between newly emerging jurisdictions that are organised along functions instead of territories. The fifth freedom suggested here allows for such *functional, overlapping competing jurisdictions*. They will be called by their acronym FOCJ (any one jurisdiction is called FOCUS). FOCJ form a federal system of governments that is not dictated from above, but emerges from below as a response to citizens' preferences. This fifth freedom requires a constitutional decision (B.S. Frey, 1983; Mueller, 1996) which ensures that the emergence of FOCJ is not blocked by existing jurisdictions such as direct competitors or higher levels of government. The European Constitution must give the most basic political units (communities) a certain degree of independence so that they can engage in forming FOCJ. Citizens must be given the right to establish FOCJ by popular referenda, and political entrepreneurs must be supported and controlled by the institution of popular initiatives. The FOCJ themselves must have the right to levy taxes to finance the public services they provide.

The first section specifies the concept of FOCJ and puts it into theoretical perspective. The next section evaluates FOCJ and discusses how the problems related to this type of jurisdictions may be overcome. The third section points to historical precursors, and the following section presents contemporary examples of FOCJ. In the fifth section the concept of competitive federalism is contrasted to other proposals. The final section offers concluding remarks.

The proposal: FOCJ

The federal units proposed here have four essential characteristics. They are as follows.

- *Functional (F)*: the new political units extend over areas defined by the tasks to be fulfilled;
- *Overlapping (O)*: in line with the many different tasks (functions) there are corresponding governmental units extending over different geographical areas;
- *Competing (C)*: individuals and/or communities may choose the governmental unit they wish to belong to, and they have the political right to express their preferences directly via initiatives and referenda;
- *Jurisdictions (J)*: the units established are governmental, they have powers of enforcement and can, in particular, levy taxes.

FOCJ are based on theoretical propositions advanced in the economic theory of federalism. They nevertheless form a governmental system completely different to the one suggested in that literature. While the economic theory of federalism

analyses the behaviour of given political units at the different levels of government, FOCJ emerge in response to the 'geography of problems'.[1] The four elements of FOCJ are now related to economic theory as well as to existing federal institutions, pointing out both similarities and differences with existing concepts.

Functions

A particular public service, which benefits a certain geographical area, should be financed by the people living in that locality; there should be no spillovers. Different governmental units can cater for regional differences in the populations' preferences or, more precisely, to its demands. To minimise cost, these units have to exploit economies of scale in provision. As the latter may strongly differ between functions (say, between schools, police, hospitals, power plants and defence), there is an additional reason for uni-functional (or few-functional) governmental units of different sizes. This is the central idea of 'fiscal equivalence' as proposed by Olson (1969) and Oates (1972). This endogeneity of the size of governmental units constitutes an essential part of FOCJ. However, fiscal equivalence theory has little concern with decision-making within functional units. The supply process is either unspecified, or it is assumed that the mobility of persons (and of firms, a fact rarely mentioned) automatically induces these units to cater to individual preferences.

Overlaps

FOCJ may overlap in two respects: (i) FOCJ catering to different functions may overlap; (ii) two or more FOCJ for the same function may geographically intersect (a multitude of school FOCJ may exist in the same geographical area). An individual or a political community normally belongs to various FOCJ at the same time. FOCJ need not be physically contiguous, nor have a monopoly over a certain area. Thus, this concept completely differs from archaic nationalism with its struggle over territory. It also breaks with the notion of federalist theory that states that units at the same level should not overlap. On the other hand, in that respect it is closer to Buchanan's (1965) 'clubs' which may intersect.

Competition

The heads of FOCJ are induced to conform closely to their members' preferences via two mechanisms: the individuals' and communities' option to exit, mimicking market competition (Hirschman, 1970), and their right to vote, establishing political competition (Mueller, 1989). It should be noted that migration is only one means of exit; often, membership in a particular FOCUS can be discontinued without changing one's location. Exit is not restricted to individuals or firms; as indicated earlier, political communities as a whole, or parts of them may also exercise this option. Moreover, exit may be total or only partial. In the latter case, an individual or community only participates in a restricted set of FOCUS activities.

Secession rights ought to be an important ingredient for a future European constitution, as advocated by Buchanan (1991) and the European Constitutional

Group (1993). This is in marked contrast to the prevailing system within nation states and federations where the right to secede is usually absent and prevented by force. Current European treaties do not provide for the secession of a nation from the European Union, and *a fortiori* for part of a nation.

For FOCJ to establish competition between governments, exit should be as unrestrained as possible. In contrast, entry need not necessarily be free. As in Buchanan-type clubs, jurisdictions and individuals may be asked to pay a price if they want to join a particular FOCUS and benefit from its public goods. The existing members of the particular FOCUS have to democratically decide on an adequate entry price for a new member.

Competition also needs to be fostered by political institutions, as the exit option does not suffice to induce governments to act efficiently. Citizens should directly elect the persons managing the FOCJ, and should be given the right to initiate popular referenda on specific issues. These democratic institutions are known to raise efficiency in the sense of catering to individual preferences; on elections, see Downs (1957) and Mueller (1989); on referenda B.S. Frey (1994).

Jurisdictions

A FOCUS is a democratic governmental unit with authority over its citizens, including the power to tax. Two forms of membership can be distinguished. First, all citizens automatically become members of the FOCJ to which their community belongs. In that case, an individual can only exit via mobility. Second, individuals may freely choose whether they want to belong to a particular FOCUS, but while they are its citizens, they are subject to its authority. Such FOCJ may be involuntary in the sense that one must belong to a FOCUS providing a certain function, say schooling, and must pay the corresponding taxes. An analogy here is health insurance, which in many countries is obligatory, but where individuals are allowed to choose an insurance company. The citizens of such a school-FOCUS may then decide that everyone must pay taxes in order to finance a particular school, irrespective of whether one has children. With respect to FOCJ providing functions with significant redistributive effects, a minimal regulation by the central government may be in order so that, for instance, citizens without children do not join 'school-FOCJ' which in effect do not offer any schooling and have correspondingly low (or zero) taxes. In this respect, Buchanan-type clubs differ from FOCJ, because they are always voluntary, while membership in a FOCUS can be obligatory.

An evaluation of FOCJ

Benefits

FOCJ compare favourably to traditional forms of federalism. Due to the concentration on one functional area, the citizens of a particular FOCUS have better information on its activity, and are in a better position to compare its performance with other governments. As many benefits and costs extend over a relatively limited

geographic area, FOCJ are often likely to be small. The exit option, which arises from the existence of overlapping jurisdictions, can guarantee that suppliers take individual preferences into account.

On the other hand, FOCJ are able to provide public services at low cost because they are formed in order to minimise interjurisdictional spillovers and to exploit economies of scale. When the benefits of a specific activity indivisibly extend over large areas, and there are decreasing costs, the corresponding optimal FOCUS may cover many communities, several nations, or even Europe as a whole. An example may be defence against outward aggression where the appropriate FOCUS may extend over the whole of Europe (and beyond the European Union).

The threat of exit by dissatisfied citizens or communities, and the benefit of new citizens and communities joining the FOCUS, gives an incentive to take individual preferences into account and to provide the public services efficiently. Another advantage of FOCJ is that they open up the politicians' cartel (*classe politique*) to functionally competent outsiders. While all-purpose jurisdictions attract persons with non-specialised knowledge into the political arena, in FOCJ, persons with a well-grounded knowledge in a particular functional area (say education or refuse collection) are successful.

A federal web composed of FOCJ certainly affects the role of nation states. They will certainly lose functions they presently do not fulfil according to the population's preferences, or which they produce at higher cost than FOCJ designed to exploit cost advantages. On the other hand, the scheme does not purport to do away with nations but allows for multi-national as well as small-scale alternatives where citizens desire them. Nation states continue in so far as they provide functions efficiently according to voter preferences.

Alleged problems

Up to this point the advantages of FOCJ have been emphasised. However, there are also possible problems that are now discussed.

Overburdened citizens

In a federal system of FOCJ, each individual is a citizen of various jurisdictions. As a consequence, individuals may be overburdened by voting in elections and referenda taking place in each FOCUS. However, citizens in a direct-democratic FOCUS find it much easier to politically participate as they have only to assess one or a few concrete issues at a time.

Overburdened consumers

An individual is confronted with a multitude of suppliers of public services, which could make life difficult. This is the logical consequence of having more options to choose from, and is similar to supply in the private sector. If citizens neverthe-

less find it to be a problem, a governmental or a private advisory service can be established which offers information and support to consumers.

'Need' to coordinate the activities of FOCJ

While co-ordination is obviously often needed, co-ordination between governments is not necessarily beneficial. It sometimes serves to build cartels among the members of the *classe politique* who then evade or even exploit the population's wishes (see CEPR, 1993; Vaubel, 1994; B.S. Frey, 1994). As far as welfare increasing co-ordination is concerned, its need is reduced because the FOCJ emerge in order to minimise externalities. If major spillovers between FOCJ exist, new FOCJ will be founded taking care of these externalities.

Redistribution

It may be claimed that all forms of federalism – including FOCJ – undermine redistributive goals. Moreover, the formation of FOCJ could be based on members' income. This fear is unwarranted, as long as redistribution is based on solidarity, or on insurance principles. A problem emerges when redistribution is a pure public good, and requires enforcment to prevent free-riding. However, recent empirical research (Gold, 1991; Kirchgässner and Pommerehne, 1996) suggests that substantial redistribution is feasible in federal systems.

Historical precursors

Decentralised, overlapping political units have been an important feature of European history. The competition between governments in the Holy Roman Empire of German Nations, especially in what is now Italy and Germany, has been intensive. Several of these governments were small. Many scholars attribute the rise of Europe to this diversity and competition between governmental units, which fostered technical, economic and artistic innovation (see Hayek, 1960; Jones, 1981; Rosenberg and Birdzell, 1986; and Weede, 1993). The unification of Italy and Germany in the nineteenth century, which has often been praised as a major advance, partially ended the stimulating competition between governments and led to deadly struggles between nation states. Some smaller states escaped unification: Liechtenstein, Luxembourg, Monaco, San Marino and Switzerland stayed politically independent, and at the same time grew rich.

The above-mentioned governmental units were not FOCJ in the sense outlined in this contribution but they shared the characteristic of competing for labour and capital (including artistic capital) among each other. However, history also reveals examples of jurisdictions even closer to FOCJ. The problems connected with Poland's strong ethnic and religious diversity (Catholics, Protestants and Jews) were at least partly overcome by jurisdictions organised along these lines, and not along geography (Rhode, 1960). The highly successful Hanse prospered from the twelfth to the sixteenth centuries, and comprised *inter alia* Lübeck, Bremen, Köln

(today German), Stettin and Danzig (today Polish), Kaliningrad (today Russian), Riga, Reval and Dorpat (today parts of the Baltic republics) and Groningen and Deventer (today Dutch). Furthermore, London (England), Bruges and Antwerp (today Belgian) and Novgorod (today Russian) were Handelskontore or associated members. It was clearly a functioning governmental unit providing trade rules and facilities that were not geographically contiguous.

Partial existence today

In two countries functional, overlapping and competing jurisdictions exist to some degree. They do not in all cases meet all the requirements of FOCJ specified above, but nevertheless show that democratic functional jurisdictions are viable.

US special districts

Single-purpose governments play a significant role in the American federalist system. Their number has increased more quickly than other types of jurisdictions (Zax, 1988). There are both autonomous and democratically organised as well as dependent special districts (say for fire prevention, recreation and parks). Empirical research suggests that the former type is significantly more efficient (Mehay, 1984). Existing jurisdictions tend to oppose the formation of special districts. In order not to threaten the monopoly power of existing municipalities, statutes in eighteen states prohibit new municipalities forming within a specified distance from existing municipalities; in various states there is a minimum population requirement and various other administrative restrictions have been introduced (Nelson, 1990). Empirical studies reveal that these barriers tend to reduce the relative efficiency of the local administration (DiLorenzo, 1981; Deno and Mehay, 1985), and tend to push local government upwards (Martin and Wagner, 1978).

Swiss communes

Many Swiss cantons have a structure of overlapping and competing functional jurisdictions that share many features of FOCJ. For example, in Zürich (with a population of 1.2 million, an area of 1,700 square kilometres, and tax revenue of CHF 2,800 million) there are 171 political communes (with a tax revenue of CHF 3,900 million) which in themselves are composed of three to six independently managed, democratically organised communes devoted to specific functions and raising their own taxes. Examples for such types of functional communes can be found in Zürich as well as in the cantons of Glarus and Thurgau (for the latter, see Casella and B.S. Frey, 1992). Cantonal bureaucracy and politicians have made various efforts to suppress this diversity of functional communes. However, most of these attempts have been thwarted as the population is satisfied with existing provision. The example from Switzerland – generally considered to be a well organised and administered country – shows that a multiplicity of functional jurisdictions under democratic control is not wishful thinking but has worked well in reality.

Comparison to other proposals

FOCJ differ in many crucial respects from other proposals for a future European constitution. One of the most prominent is Buchanan (1991) who stresses individual nation's right to secede but, somewhat surprisingly, does not build on Buchanan-type clubs. The European Constitutional Group (1993) focuses on the example of the American constitution, and presents constructivist proposals with respect to the houses of parliament and the apportioning of weighted voting rights amongst countries. Overlapping jurisdictions and referenda are not considered, and the exit option is strongly restricted. Other economists (such as Blöchliger and R.L. Frey, 1992; Schneider, 1992) suggest a strengthening of federalism in the traditional sense (i.e. with multi-purpose federal units) but do not envisage overlapping jurisdictions. Reports by the Centre for Economic Policy Research (CEPR, 1993, 1995) criticise 'subsidiarity' (as used in the Maastricht Treaty) as an empty concept arguing that good theoretical reasons must be provided for central government intervention. But the report does not deal with the institutions necessary to guarantee that policy follows such theoretical advice. The idea of overlapping, not geographically based jurisdictions is briefly raised (CEPR, 1993: 54–5), but is not institutionally or practically worked out, nor is the need for a democratic organisation and the power to tax recognised.

The proposal by politicians at the European level (Herman report of the European Parliament, 1994) mainly deals with the organisation of the parliamentary system and national voting rights, and to a substantial extent accepts the existing treaties as the founding blocks of the European constitution. The crucial idea of competition between governments is neglected; the report prefers to speak of the necessary 'co-operation' between governments – which in actual fact often serves to undermine the threat of competition.

FOCJ are also quite different from the regions envisaged in existing European treaties and institutions (see Adonis and Jones, 1991). A major difference is that FOCJ emerge from below while the 'European regions' tend to be established from above. Moreover, their existence strongly depends on the subsidies flowing from the European Union and the nation states (Sharpe, 1993). In contrast, the concept of FOCJ corresponds to Hayek's (1960) non-constructivist process view. One cannot, *a priori*, determine which FOCJ will be efficient in the future. This must be left to the competitive democratic process at the level of individuals and communities. The central European constitution must only make sure that no other governmental units, particularly nations, obstruct the emergence of FOCJ. In contrast to Hayek, however, the scheme allows for a (closely restricted) set of central regulations, as mentioned above. Moreover, Hayek measures efficiency by survival in the evolutionary process while efficiency is defined here in terms of the fulfillment of citizens' demands.

'Subsidiarity' as proclaimed in the Maastricht Treaty is generally recognised to be more a vague goal than a concept with content (see EPR, 1993: 19–23). Even if subsidiarity were taken seriously, it would not lead to a real federal structure because many (actual or prospective) members of the European Union are essentially

unitary states without federal sub-units of significant competence (examples are the Netherlands, France or Sweden). The 'regions' existing in the European Union (examples are Galicia and Cataluña in Spain, or South Tyrol and Sicily in Italy) are far from being units with significant autonomous functional and fiscal competencies.

The Council of Ministers is a European decision-making institution based on federal principles (but only nations are represented) and organised according to functional principles (or at least according to the corresponding administrative units). However, this Council is only indirectly democratic (the ministers are members of governments which are democratically legitimised by the representative system) and the deliberations are not public. Exit from the European Union is not formally regulated, and exceptions to specific aspects of agreements reached (as in the Maastricht Treaty concerning the European Monetary Union, the Protocol on Social Policy, or in the Schengen Treaty concerning the free movement of persons) are granted reluctantly. Indeed, they are seen as damaging the 'spirit of Europe'. Whether differential degrees of European integration are framed as models of 'variable geometry', 'multi-track', 'multi-speed', 'two-tier', 'hard core', 'concentric circles', or as 'Europe à la carte', it always elicits fierce opposition. In a system of FOCJ, in contrast, functional units not covering everyone are taken as a welcome expression of heterogeneous demands among European citizens.

Concluding remarks

In view of the major advantages of FOCJ the economist's standard question arises: if this type of federalism is so good, why is it not more in evidence?

The organisation of states today does not follow the model of FOCJ for two major reasons. An obvious, but crucial one, is that individuals and communities are prohibited from establishing such jurisdictions, and in many countries of the European Union, communities are not even allowed to formally collaborate with each other without the consent of the central government (see Sharpe, 1993, esp. 123ff.).

Second, a system of FOCJ cannot be observed because it violates the interests of politicians and public officials at the higher levels of government. The emergence of FOCJ reduces the public supplier's power and increases citizens' influence via newly introduced mechanisms of competition by exit and entry, and by elements of direct democracy. Both are regularly opposed by the *classe politique*. As politicians' discretionary room and therefore the rents appropriable are the larger, the higher the federal level, they favour a shift of competencies in that direction, and oppose local decision-making, especially by FOCJ, wherever possible.

In the countries of the European Union (and elsewhere) a federal system of FOCJ will not arise if these barriers are not overcome. A necessary condition are new constitutional rules allowing the formation of FOCJ and giving citizens and governments the right to appeal to the Constitutional Court in case they are blocked.

FOCJ are a flexible concept. They can be introduced in small doses. An obvious first application would be functional units straddling communes situated on two or more member states of the European Union. Such FOCJ would make a substantial contribution to the coming together of Europe at a level directly benefiting the citizens. Such jurisdictions would contribute more to the emergence of a 'European spirit' than grandiose pronouncements by heads of states or sumptuous conferences and meetings.

Another obvious scope for applying FOCJ is the integration of the countries of the ex-Soviet bloc into Europe. At present, the European Union insists that these nations fully accept the '*acquis communautaire*' though their economic and institutional development differs drastically from those of the present member states. Even staunch supporters of the present European system have to accept that it is impossible to integrate these countries into the EU without changing its constitution. This would present an excellent opportunity to open up the EU constitution to overcome the democratic and decentralisation deficits.

Acknowledgement

The ideas developed here are based on joint research with my co-worker Reiner Eichenberger, see Frey and Eichenberger (1995, 1996, 1999).

Note

1 Similar ideas can be found in by Montesquieu (1749). Burnheim (1985) mentions similar elements. In the economics literature a related concept has been pioneered by Tullock (1994), who calls it 'sociological federalism'. Casella and Frey (1992) discuss the concept and refer to relevant literature.

References

Adonis, A. and S. Jones (1991) 'Subsidiarity and the European Community's constitutional future', *Staatswissenschaft und Staatspraxis* 2 (2): 179–96.

Bird, R. M. (1993). 'Threading the fiscal labyrinth: some issues in fiscal decentralization', *National Tax Journal* 46: 201–21.

Blöchlinger, H. and R.L. Frey (1992) 'Der schweizerische Föderalismus: Ein Modell für den institutionellen Aufbau der Europäischen Union', *Aussenwirtschaft* 47: 515–48.

Breton, A. (1996) *Competitive Governments: An Economic Theory of Politics and Public Choice Finance*, New York: Cambridge University Press.

Buchanan, J.M. (1965) 'An economic theory of clubs', *Economica* 32 (February): 1–14.

Buchanan, J.M. (1991) 'An American perspective on Europe's constitutional opportunity', *Cato Journal* 10 (3): 619–29.

Burnheim, J. (1985) *Is Democracy Possible?: The Alternative to Electoral Politics*, Cambridge: Polity Press.

Casella, A. and B.S. Frey (1992) 'Federalism and clubs: towards an economic theory of overlapping political jurisdictions', *European Economic Review* 36: 639–46.

CEPR (Centre for Economic Policy Research) (1993) *Making Sense of Subsidiarity: How Much Centralization for Europe?* London: CEPR.

CEPR (Centre for Economic Policy Research) (1995) *Flexible Integration. Towards a More Effective and Democratic Europe*, London: CEPR.

Deno, K.T. and S.L. Mehay (1985) 'Institutional constraints on local jurisdiction', *Public Finance Quarterly* 13: 450–63.
DiLorenzo, T.J. (1981) 'Special districts and local public services', *Public Finance Quarterly* 9: 353–67.
Downs, A. (1957) *An Economic Theory of Democracy*, New York: Harper and Row.
European Constitutional Group (1993) *A European Constitutional Settlement*, (Draft) London.
Frey, B.S. (1983) *Democratic Economic Policy*, Oxford: Blackwell.
Frey, B.S. (1994) 'Direct democracy: politico-economic lessons from Swiss experience', *American Economic Review* 84 (2): 338–48.
Frey, B.S. and R. Eichenberger (1995) 'Competition among jurisdictions. the idea of FOCJ', in Lüder Gerken (ed.) *Competition among Institutions*, London: Macmillan.
Frey, B.S. and R. Eichenberger (1996) 'FOCJ: competitive governments for Europe', *International Review of Public Economics* 16: 315–27.
Frey, B.S. and R. Eichenberger (1999) *The New Democratic Federalism for Europe: Functional Overlapping and Competing Jurisdictions*, Cheltenham: Edward Elgar.
Gold, S.D. (1991) 'Interstate competition and state personal income-tax policy in the 1980s', in D.A. Kenyon and J. Kincaid (eds) *Competition among States and Local Governments*, Washington DC: Urban Institute Press.
Hayek, F.A. von (1960) *The Constitution of Liberty*, London: Routledge.
Herman, F. (Reporter) (1994) *Zweiter Bericht des Institutionellen Ausschusses über die Verfassung der Europäischen Union*, Sitzungsdokumente A3–0064/94, Europäisches Dokument.
Hirschman, A.O. (1970) *Exit, Voice and Loyality*, Cambridge, MA: Harvard University Press.
Jones, E.L. (1981) *The European Miracle*, Cambridge: Cambridge University Press.
Kirchgässner, G. and W.W. Pommerehne (1996) 'Tax harmonization and tax competition in the European Community: lessons from Switzerland', *Journal of Public Economics* 60: 351–71.
Martin, D. and R. Wagner (1978) 'The institutional framework for municipal incorporation', *Journal of Law and Economics* 21: 409–25.
Mehay, S.L. (1984) 'The effect of governmental structure on special district expenditures', *Public Choice* 444: 339–48
Montesquieu, C.L. (1749) *De l'esprit des lois*, Paris: Garnier.
Mueller, D.C. (1989) *Public Choice II*, Cambridge: Cambridge University Press.
Mueller, D.C. (1996) *Constitutional Democracy*, New York: Oxford University Press.
Nelson, M.A. (1990) 'Decentralization of the subnational public sector: an empirical analysis of the determinants of local government structure in metropolitan areas in the U.S', *Southern Economic Journal* 57: 443–57
Oates, W.E. (1972) *Fiscal Federalism*, New York: Harcourt Brace Jovanovich.
Olson, M. (1969) 'The principle of 'fiscal equivalence': the division of responsibilities among different levels of government', *American Economic Review* 59 (2): 479–87.
Rhode, G. (1960) 'Staaten-Union und Adelsstaat: Zur Entwicklung von Staatsdenken und Staatsgestaltung in Osteuropa, vor allem in Polen/Litauen, im 16. Jahrhundert', *Zeitschrift für Ostforschung* 9: 185–215.
Rosenberg, N. and L.E. Birdzell (1986) *How the West Grew Rich: The Economic Transformation of the Industrial World*, London: I.B. Tauris.
Schneider, F. (1992) 'The federal and fiscal structures of representative and direct democracies as models for a European federal union: some ideas using the public choice approach', *Journal des Economistes et des Etudes Humaines* 3: 403–37.
Sharpe, L.J. (ed.) (1993) *The Rise of Modern Government in Europe*, London: Sage.
Tullock, G. (1994) *The New Federalist*, Vancouver: Fraser Institute.
Vaubel, R. (1994) 'The political economy of centralization and the European Community', *Public Choice* 81: 151–90.

Weede, E. (1993) 'The impact of interstate conflict on revolutionary change and individual freedom', *Kyklos* 46: 473–95.
Zax, J.S. (1988) 'The effects of jurisdiction types and numbers on local public finance' in Harvey S. Rosen (ed.) *Fiscal Federalism: Quantitative Studies*, Chicago: The University of Chicago: 79–106.

3 On the economic causes of contemporary civil wars

Tony Addison and S. Mansoob Murshed

Development economists have traditionally discussed the design of policy independently of conflict and its occurrence, these being seen as issues for political scientists. But the implications for social conflict of economic decisions cannot be ignored in this way. Similarly, the potential for conflict and civil war in retarding growth and development are equally important. The potential for disaster in getting economic policy wrong is high in fragile low-income societies that need economic growth, especially in those whose resources are easily capturable, giving them an above-average vulnerability to conflict.

All societies are characterised by some degree of conflict over political and civil rights, employment opportunities, and access to social and economic services. These conflicts may be based on social class, ethnicity, religion, region or some combination of these factors. In stable societies, conflict is channelled into non-violent means for both its expression and resolution. Since rapid economic growth is impossible without a large measure of social stability, investment in institutions for non-violent conflict resolution is as important to development as investment in physical or human capital. Today's rich countries built over time formal and informal institutions of conflict management – thereby damping down often deep and bitter hatreds. Fundamentally, violent internal conflict implies the absence or breakdown of the implicit or explicit social contract.

It is important to understand that violence is an alternative to peaceful production as a form of economic activity. The work of Francis Edgeworth, writing in the late nineteenth century, provides a useful starting point in understanding the economic rationale for violence. Edgeworth distinguished between consent – and its absence – in human economic interaction:

> The first principle of Economics is that every agent is actuated only by self-interest. The workings of this principle may be viewed under two aspects, according as the agent acts *without*, or *with*, the consent of others affected by his actions. In wide senses, the first species of action may be called *war*; the second, contract.
>
> (Edgeworth, 1881: 16–17)

Or, as Garfinkel and Skaperdas (2000) put it: in securing an income, humanity has a choice between production and predation, the relative returns being in part

determined by the cost of 'swords' relative to 'ploughshares'. The institutional environment, the quality of law and contract enforcement also determine this choice (see Mehlum *et al.* 2001). Criminal activity, whether taking the form of extortion or theft, is only one aspect of the economics of violence. War, especially civil war, also has an economic dimension. Hirshliefer (1995) models anarchic inter-group warfare using non-cooperative game theory, in a setting reminiscent of the primitive conflict over resources between neighbouring communities. Most contemporary wars are civil wars taking place within the nation state, and nearly all present-day warfare is outside the developed world (Wallensteen and Sollenberg, 2000).

An important question is whether contemporary civil wars are a simple product of grievances or the result of criminal acquisitive desire. Genuine grievances can produce civil war, and we elaborate on some examples below. Many of these are related to economic factors: systematic economic discrimination against specific groups based on ethno-linguistic or religious differences. Extreme poverty and poor social conditions (epitomised by refugee camps) also facilitate conflict by providing a pool of aggrieved and more readily available combatants.[1] Conflict can also be motivated by simple greed, such as the desire to control rents from natural resources such as oil or diamonds (Collier and Hoeffler, 1999; Berdal and Malone, 2000). This is more likely in the absence of a functioning social contract, and a state incapable of guaranteeing security and carrying out its functions of re-distribution. The profit to be made from Sierra Leone's conflict diamonds is aptly summed up by one courier involved in smuggling diamonds to London's jewellery centre, Hatton Garden: 'A batch of diamonds that cost £1,000 up country in Sierra Leone – that includes production costs, bribes and, if necessary, murder costs – is worth £5,000 by the time it reaches Freetown and £20,000 by the time it gets to Hatton Garden' (Lashmar, 2000: 14). Greed and grievance interact to produce conflict and reproduce themselves. Greed disguised as grievance can lead to civil wars, which results in greater grievance as people (or their loved ones) are killed or hurt. Criminal activities, whether involving extortion or trade in contraband goods, play a major part in financing conflict. Warlords profit from conflict, both directly, and by taking advantage of an absent, weak or pre-occupied state. Aggrieved diasporas living in peaceful and prosperous nations frequently contribute to financing conflicts in the countries they have fled from. External actors, such as the great powers, former colonial states and neighbouring nations may play a significant role in triggering or sustaining conflict.

The relationship between nations is also marked by conflict – over markets, natural resources, and territories. The work of Findlay (1996) elegantly characterises territorial expansion or disintegration in terms of an extended production function. International trade and commerce can promote peaceful co-existence and liberal democracy, as well as providing the motivation for violent competition (see Polachek, Robst and Chang, 1999). At some critically high level of economic interdependence countries will cooperate; trade with one another rather than fight. Liberal democracy may promote this process. This view cannot, however, explain secession and civil war as regions and ethnic groups within a nation state have highly interdependent economies in terms of the exchange of goods, services and factors of production. Hegre *et al.* (2001) have demonstrated a U-shaped relation

between democratic institutions and the incidence of civil war over time. The probability of civil conflict is lowest both in established, well-functioning democracies, and perfect autocracies. It is at some intermediate stage between autocracy and democracy that the risk of internal conflict is greatest. Note that highly-developed autocracies and democracies both imply the absence of state failure.

Successful economic development is dependent upon effective mechanisms for the peaceful resolution of conflict between states and within the national level. Thus, while Eritrea has been quite successful in building a cohesive nation, Eritrea and Ethiopia have failed to peacefully settle their differences over trade and territory (Addison, 2000a). Likewise, Uganda's success in rebuilding after conflict is endangered by escalating tensions with Rwanda. In Africa, conflict has contributed greatly towards the growth collapse in the region, exacerbated by adverse initial conditions and poor policy decisions.

The 1990s witnessed the disintegration of the former Yugoslavia, genocide in Rwanda, and the collapse of the Somali state, to name just three of the 43 major conflicts that occurred in that decade.[2] But such horror is not unique to our time – the 1970s saw genocide in Cambodia – and while the number of conflicts rose over 1990–95, ethnic warfare at least started to decline after 1995, leading Ted Gurr to conclude that by '. . . the late 1990s, the most common strategy among ethnic groups was not armed conflict but prosaic politics' (Gurr, 2000: 53). Nevertheless, ethnic or religious grouping continues to provide a more powerful basis for collective action compared to social class, for example. In addition, a history of injustice against an identifiable ethnic group can serve as both an organising principle and a justification for the violent appropriation of the proceeds of natural resources, particularly diamonds, but also oil, timber, ivory, and narcotics. A wide variety of methodologies exist for the study of conflict and civil war.

Our focus is on work in economics and rational-choice political science, both of which see conflict as amenable to analysis using choice-theoretic behaviour. This chapter is organised as follows. The first section sets out the economic causes of conflict, focusing on relative deprivation versus the contest over natural resource wealth. It also discusses the intensity of war and its financing. The second part turns to the problem of getting credible commitments to peace, especially when belligerents have poor reputations. The third section focuses on the difficulties of reconstruction, and the opportunities for designing recovery so that the social contract is rebuilt and peace is secured. The final section concludes by summarising the main arguments and highlighting the importance of conflict-prevention to economic development.

The origins of civil war and its persistence

Understanding the motivation for conflict is clearly central to an understanding of why some countries avoid conflict while others fail. To organise our discussion we first discuss the role of relative deprivation and then the contest over resource wealth. We then focus on issues of financing a civil war and its intensity.

Relative deprivation

Relative deprivation – the perception by one or more parties that they are unjustly treated – is a major cause of civil war. Many conflict societies are characterised by large inequalities in access to the productive assets necessary for livelihoods and in public spending on economic and social infrastructure and services. Research on conflict has emphasised the importance of horizontal inequalities (by ethnicity, socio-economic class, etc) as sources of conflict (Nafziger *et al.*, 2000, Klugman, 1999 and Stewart, 2000). Here, we note three dimensions of the problem:

Discrimination in public spending and taxation

Managing the finances of the state is one of the most difficult aspects of development – few countries do it well. Some countries do it so badly that discrimination in the allocation of public spending, and unfair tax burdens, lead to serious unrest. Grossman (1991) develops a theoretical model of insurrection against the state by the peasantry reacting to over taxation, where the state is a tax-farmer interested in maximising the income of the rentier class. Discrimination in the allocation of public employment is particularly resented in societies in which public employment represents the principal avenue for personal advance. In addition, the over-taxation of smallholders encourages insurrection, and indigenous peoples often face discrimination in access to schooling, health care, and public-sector jobs, a factor in Guatemala's thirty-year civil war, for example. Both Burundi and Rwanda have suffered from highly distorted structures of public spending, which discriminated against disadvantaged ethnic groups to the benefit of political elites who controlled spending. Public expenditure and taxation reforms to lessen inequalities are difficult to implement in institutionally weak states; wealthy groups often control the legislatures, enabling them to block progressive tax reforms such as land taxes.

Where there are inter-group fiscal transfers, which may take the form of spending on education and health for disadvantaged groups, or including them in government employment, commitment to the transfer by those in power may be imperfect. This lack of credibility of the implicit or explicit transfer can eventually lead to civil war (Azam and Mesnard, 2001). On the other hand, societies such as Malaysia have managed a federal system of taxes and benefits relatively equitably and efficiently so as to avoid civil war for over three decades.

High asset inequality

Agrarian societies with high income inequality – for example El Salvador, Guatemala, the Philippines, and Zimbabwe – have high asset inequality, and are very prone to conflict. In these societies, agrarian elites use their collateral to further leverage their existing wealth through a financial system that they control by means of family/business cross-holdings. The coffee economies of Central America – which have had recurrent conflicts – display this nexus *par excellence* (Paige, 1997). Land inequality, the dispossession of peasant communities, and the limited

poverty reduction associated with economic growth in highly unequal societies provide fertile ground for insurrection – the Chiapas rebellion in Mexico is one example. The repudiation of the results of free and fair elections in Fiji is in part the result of long-running conflict over land, thus demonstrating that investment in political processes is insufficient for peace if underlying tensions over asset inequalities are not satisfactorily addressed.

Asset redistribution such as land reform to lessen inequality is more difficult than public finance reform. Thus, programmes supported by the World Bank in Guatemala aim to redress discrimination in public spending against indigenous peoples, thereby strengthening both poverty reduction and social peace. At the same time resolving land issues in Guatemala appears to be going nowhere. Similarly, in South Africa, more progress is apparent in redressing racial imbalances in the allocation of public spending than in access to land. While this is a positive step, the situation remains fragile in the presence of economically-disenfranchised groups in South Africa.

Economic mismanagement and recession

In Africa, Latin America and the former Soviet Union conflict ridden countries have also suffered prolonged economic mismanagement and growth collapse. Successive IMF and World Bank supported adjustment programmes in DRC-Zaire, Somalia, Russia and elsewhere not only proved incapable of promoting economic recovery, but given the level of corruption within the state, themselves became targets to be captured by elite groups. Corruption and rent seeking within the state is closely associated with the collapse of growth, see Murshed and Perälä (2001), and the outbreak of violent conflict. Azam (2001) explains inter-ethnic conflict in Africa in terms of state failure. The failure of the state to meet its obligations regarding public goods provision and security, forces citizens to fall back on more reliable ethnic ties or ethnic capital akin to social capital. This reliance, in turn, can encourage outright warfare with the state, or conflict between different ethnic groups over economic resources including the revenues from criminal activities. Such an explanation of civil conflict is not only valid in Africa, but also in the successor states to the Soviet Union, where the state has repudiated an earlier social contract, or in Latin America where revolutionary groups and drug-barons are more reliable providers of public services than a weak, corrupt or indifferent state.

Economic mismanagement is often associated with an uneven and unfair distribution of the burdens of subsequent adjustment; public spending benefiting the elite and the military is protected (often favouring particular ethnic groups), with the burden of adjustment placed on expenditures of value to the poor and disadvantaged groups. This was a factor in the build up to the recent conflicts in Burundi, Guinea-Bissau, and Rwanda. Moreover, countries have very different tolerances with regard to internal and external shocks. Rodrik (1998) demonstrates that many of the largest falls in output following shocks occur in divided societies (as measured by variables such as income inequality and ethnic fragmentation) with weak institutions of conflict management, including the rule of law and democracy.

These societies are unable to equitably share the burdens of adjustment (or protect the poor), and social conflict magnifies the economic impact of the shock, leading to growth collapses. This may be an apt description of events in Indonesia over the last few years; certainly Indonesia has weathered the Asian crisis much less successfully than either Malaysia or Thailand.

In summary, the policy prescription that broad-based development – that is development aimed at raising living standards for all, not just the privileged – is good for peace is a sound one. But, there are further complications that arise from the contest over valuable natural resource wealth, and it is to these that we now turn.

The contest over natural resource wealth

In a number of cases war is closely associated with natural resource exploitation (see Table 3.1). Collier and Hoeffler (1999) find empirical evidence showing that a relatively high dependence on primary commodity exports is strongly correlated, in a non-monotonic way, with the occurrence of war. Natural resources constitute 'booty'. In many cases this booty is far more valuable than possession of the state itself – with its assets and powers of taxation – and belligerents may feel no regret in destroying state institutions if they can directly control and mine the country's natural resources. Thus, belligerents in the wars of natural-resource rich countries act in ways that are closer to what Mancur Olson (1996) called 'roving bandits' – who have no encompassing interest in preserving the state or its people but are simply intent on loot – than to 'stationary' bandits who take control of the state and seek to maximise their own profit by encouraging stability and growth in their new domain.

Civil wars motivated by the desire to control natural resource rents are akin to 'warlord competition', a term that owes its origins to the violent competition between leaders in the context of medieval European cities (Skaperdas, 2002). Competing warlords have to satisfy the participation constraints of their henchmen, but offer the citizenry a semblance of security in return for a tribute. Excessive competition between these warlords means intense warfare, and a reduction of the very economic surplus which each warlord aims to milk. A similar analysis can be applied to gang warfare and the market for extortions (Mehlum *et al.* 2001). The important point is why warlords do not cooperatively divide the surplus, instead of fighting each other, which only serves to reduce the attainable spoils. This apparent lack of rationality can only be explained by misperception, the absence of contract enforcement, deep-seated hatred or an emphasis on short-term individual gain.

Addison *et al*. (2000) construct a game-theoretic model of contemporary conflict involving the competition for resources. In addition to resource rents, grievances also play their part in fuelling conflict by explaining inter-group non-cooperation and serving to lower the cost of participation in conflict. Additionally, they distinguish between two main types of resource exploitation: *point resources*, which mostly involve the *extraction* of non-renewable resources (minerals), require little labour input and are geographically concentrated; and *diffuse resources*, such

Table 3.1 Conflicts in Africa and the role of natural resources

Country	*Date*	*Deaths*	*Point resources*	*Diffuse resources*
Algeria	1992–	70,000	oil, gas	
Angola	1975–	500,000	oil, diamonds	timber, ivory
Cameroon/Nigeria	1997	<1,000	oil	
Chad	1980–94	300,000	oil, uranium	
Congo-Brazzaville	1993, 1997	9,000	oil	
Democratic Republic of the Congo / Zaire	1993–	200,000	copper, cobalt, diamonds, gold	timber
Kenya	1991–	2,000		cattle
Liberia	1989–96	175,000	iron, diamonds, rubber	timber, drugs
Mozambique	1976–95	1,000,000		shrimps, ivory, timber
Rwanda	1990–	650,000		coffee
Senegal	1997–	<1,000		drugs
Sierra Leone	1991–1999	80,000	diamonds, rutile bauxite	timber
Somalia	1988–	n.a.		bananas, camels
South Africa	1990s	200,000		drugs
Sudan	1983–	1,600,000	oil,	cattle, timber
Western Sahara	1976–	n.a.	phosphates	

Source: Balencie and de La Grange (1999).

Note: The number of deaths in South Africa includes crime.

as those which mostly involve the *production* of renewable resources (crops), require large amounts of labour, and are spread geographically (Auty, 1998). In this regard, a high mineral endowment is correlated with a high incidence of conflict, while an abundance of renewable resources is not (de Soysa, 2000). Murshed and Perälä (2001) find that point resource endowed nations experience marked growth failure for two decades or more, attributable to their own social fragmentation.

The nature and geographical distribution of the point resources can affect the nature of the conflict itself. For example, in Congo-Brazzaville, where offshore oil is the only mineral resource, the rebels needed to capture the capital city together with the main port, which they succeeded in doing in the 1997 civil war. In contrast, Angola's two mineral resources (offshore oil and alluvial diamonds) have enabled both sides (the MPLA government, and the UNITA rebels) to engage in protracted warfare. Government and UNITA leaders enjoy considerable wealth, from oil and diamonds respectively, and in the absence of decisive victory by either side, the MPLA retains control of the capital city and the state apparatus, while UNITA controls many rural areas (Le Billon, 1999). In Liberia and Sierra Leone, the diversity of resources (rubber, timber, diamonds, and iron ore) and their geographical spread led to the phenomenon of warlordism characterised by

a highly-fragmented conflict between a weak government holding the capital and numerous armed groups controlling resources in the interior (Reno, 1998).

The intensity and financing of war

Our examples above tend to imply that belligerents prefer outright victory. But this is not always the case. For long periods, some or all belligerents may prefer a situation of war to either peace or decisive victory.[3] Thus, the military forces of both the government and the rebels may derive considerable wealth and status from the continuation of war, while decisive victory would imply their demobilisation and loss of accumulation opportunities. The result is often an avoidance of total war and instead a form of 'low-intensity' conflict that minimises direct losses to belligerents, and increases the direct gains (booty). Large battles are avoided, and at the local level belligerents may in fact cooperate for extended periods when it is mutually profitable (splitting booty, profitable smuggling across frontlines, and cooperation or non-interference in looting civilians and humanitarian agencies). This behaviour leads to a 'comfortable military stalemate' (Zartman, 1995). Angola, Cambodia, Liberia, and Sierra Leone are just a few examples of 'profitable conflicts', often characterised by periods of cooperation between belligerents.

The 'low intensity' of these conflicts does not imply a low impact on populations. On the contrary, most are characterised by extreme violence against civilians (Azam and Hoeffler, 2001). This violence is itself often economically motivated, being associated with looting, forced labour, or the expulsion of local communities to secure exclusive control over resources – thereby enabling migrant labour or international corporations to exploit resources (Keen, 1998). In such circumstances, civilians themselves may engage in violence in order to protect themselves or as a means of livelihood.

Relatedly, the financing of civil war must also be considered. Aside from the sale of point resources such as diamonds together with other loot, rebels may draw on three sources: grants and loans from local and foreign entrepreneurs, taxation of the local population, and contributions from diasporas. Laurent Kabila raised considerable sums by promising lucrative mineral concessions once he had toppled Mobutu Sese Seko in Zaire. In the war that continued after Kabila took power, competing rebel movements backed by Rwanda and Uganda, taxed local businesses. Rebellions in East Timor, Kosovo, and Sri Lanka have been financed by contributions from diasporas (as was the case in Eritrea in its 1998–2000 interstate war with Ethiopia). Additionally, during the cold war superpower assistance financed some civil wars; at least one of which still persists (Angola). On the government side, war adds to fiscal pressures, usually leading to hyperinflation. The resulting inflation tax partially compensates for the loss of direct and indirect taxes as the wartime economy and its tax base contracts (Addison and Murshed, 2001b). If the government controls natural resources, then it may mortgage these to raise foreign commercial loans: the case in Angola.

Getting credible commitments to peace

In countries that are at war, getting the opposing parties to the peace table, and then to successfully negotiate peace, is a daunting task – especially when either one or both parties has previously reneged on the peace process. It has to be remembered that many of today's civil wars contain a dimension of external intervention by neighbours and great powers in the past or present, helping to fuel the conflict. This may be said to be true of the wars in Angola, DRC, Sri Lanka and in many of the conflicts in the Caucasus. In a detailed empirical study Walter (2001) finds that negotiations to end civil wars break down more frequently than negotiations to end wars between states.[4] The conflicts in Angola and Guinea-Bissau are both examples of the problem encountered in getting each side to make credible commitments (Kovsted and Tarp, 1999; Le Billon, 1999). Overcoming the problem of credible commitment is therefore critical, especially for third parties such as the UN and the Africa Union intent on devising ways forward to end war – these are referred to as 'commitment technologies' in the literature. This problem, the inability to credibly commit, is exacerbated by repeated opportunistic interaction, which has been analysed within the framework of transactions-cost politics by Dixit (1996) among others. The solutions involve institution building, including institutions that restrain discretionary action by politicians through constitutional mechanisms. Democracy may be a vital first stage towards the process, although new democracies in particular can be hijacked by vested interests and pre-existing strong informal social networks, an endemic problem in Africa and the former Soviet Union (see Wedel, 2001).

The problem of credible commitment is particularly acute in so-called 'post-conflict' situations (Addison, 2000b, Addison and Murshed, 2001a). There is often considerable uncertainty following the end of civil war. For this reason some writers reject the term post-conflict as being highly misleading (see Crisp, 1998). Moreover, one or more parties to the conflict may repudiate the results of internationally recognised and supported elections intended to achieve a post-conflict settlement (Kumar, 1998). Angola is again a tragic example: UNITA rejected the results of the 1992 elections (which it lost) and returned to war. In some cases, warlords may use this implicit threat to win over an otherwise fearful electorate – this was basically the message that warlord Charles Taylor conveyed to Liberia's electorate in the 1997 presidential elections which he won (Ellman and Wantchekon, 2000: 518). If this occurs, then the 'post-conflict' situation may be highly unstable – as it is in Liberia and Sierra Leone.

An aspect of the commitment problem that has received insufficient attention is the very high discount rates, or the short time horizons of the parties involved (Addison and Murshed, 2001a). In situations of poverty and high uncertainty, agents strongly prefer a dollar today to a dollar tomorrow. Although the absolute value of future peace may be much higher than that of continued warfare, the present value may be much lower when the discount rate is very high. The same argument can be applied to reputation, a factor that is key to the credibility of peacemaking. Breaking an agreement destroys *future* reputation, but with a high enough discount rate

it might pay to renege. Each failure of the peace process raises the discount rates of the belligerents, thereby increasing the difficulty of making peace. Given the tarnished reputations of belligerents it is even harder to establish credible peace. The problem is particularly apparent in Africa where most indicators of political risk are substantially greater than elsewhere in the world (Table 3.2). Many African leaders have very high private discount rates (impatient to consume), and therefore some – particularly in the military – prefer to achieve personal advance and wealth through war rather than through peaceful investment and accumulation. Solutions lie in directly increasing the cost of reneging on peace agreements and devising commitment technologies through institutional innovation, particularly at the international level. Improving the quality of peacekeeping forces is an urgent need (as Sierra Leone demonstrated), as is increased commitment to bringing war criminals to trial. This in turn depends on thorough reform of the international system of governance (Nayyar, 2001).

We need to assess why some 'post-conflict' countries return to war (Angola) while others have managed to sustain peace (Mozambique). Again, economic motivations may lie at the root of the problem – Mozambique has few valuable minerals over which to fight while Angola has several – and this can be critical to determining the best responses by third parties intent on promoting peace. Thus donors could use aid as a lever to end conflict in Mozambique, while this lever is presently of no practical use in Angola, given the volumes of foreign exchange available to both belligerents from diamonds and oil respectively. There are numerous difficulties in designing conditionality, among them differences in objectives amongst donors (see Murshed and Sen, 1995). Raising the profitability of peaceful behaviour, and reducing profits of war-like behaviour through measures such as sanctions on the trade of 'blood' diamonds and freezing the international bank accounts of belligerents are essential complements to diplomatic endeavours.

Reconstruction and rebuilding the social contract

Contemporary wars generally reduce aggregate output – for example in Sri Lanka the cost of the war so far amounts to two years of annual GDP (Arunatilake *et al*. 2000). There may be periods of growth during wartime – from a point below the peacetime production frontier – but it will be biased towards services (and urban areas) and is therefore unlikely to be sufficiently poverty reducing.

To be successful, reconstruction must rebuild the social contract, correcting the flaws that had caused state failure. Otherwise conflict may reignite. Both political and economic reform are needed to change the rules of the game. These innovations are akin to the deep interventions described in Dixit (1999). To reduce the probability of conflict restarting, reconstruction must ensure sharing of resource-rents among some or all of the contending parties. Moreover, other sources of grievance – for instance discrimination in public employment – have to be addressed. Mere promises of reform are not enough. The pre-war social contract may have disintegrated because the credibility of such transfers was low in the context of general state failure. When designing reconstruction strategies, it has to borne in mind that

Table 3.2 Sub-Saharan Africa and other developing country groups: quality of governance, institutions and public services

(0–10, higher = better quality)[1]	*Quality of Bureaucracy 1984–98*	*Extent of Corruption 1984–98*	*Government Stability 1984–98*	*Ethnic Tensions 1984–98*	*Political Violence 1984–98*	*Law and Order 1984–98*	*Risk of Expropriation 1984–97*	*Risk of Contract Repudiation 1984–97*
Asian NIEs[2]	7.3	6.9	6.5	7.5	9.0	7.4	8.3	9.1
Asia	4.6	4.3	5.0	4.5	6.0	5.0	6.3	6.0
Advanced economies	8.7	8.3	6.5	8.2	8.7	8.7	7.9	9.2
Western hemisphere	4.2	4.6	4.9	7.1	5.8	4.8	5.7	6.0
Middle East and Europe	4.8	4.7	5.7	5.9	5.7	5.4	6.2	5.9
Sub-Saharan Africa	4.1	4.6	4.8	4.9	5.5	4.4	5.3	4.8
World	5.4	5.5	5.4	6.4	6.6	5.8	6.3	6.4
Sub-Saharan Africa countries: unweighted grouping by yearly per capita GDP growth rate in 1970–98								
High growth (top quintile)	4.3	5.9	5.5	5.7	6.4	5.1	5.6	5.3
Medium group	4.6	4.7	5.0	5.1	5.7	4.6	5.7	5.2
Low growth (bottom quintile)	3.0	3.5	3.8	3.9	4.5	3.6	4.3	3.4
Memorandum								
CFA: unweighted average[3]	4.5	4.3	4.9	5.4	6.0	4.4	5.7	5.3

Source: Reproduced from IMF (1999). Original data from Political Risk Services (1999) *International Country Risk Guide,* East Syracuse NY: Political Risk Services.

Notes

1 For regional groupings: unweighted averages of countries in the dataset.
2 Hong Kong, Korea, Singapore, and Taiwan Province of China.
3 Communauté Financière Africaine and Coopération Financière en Afrique.

state failure is more prevalent in nations with artificially designed borders based on colonial legacies (Azam, 2001, discusses this in the context of Africa). Democratisation may improve the credibility of the fiscal transfer, however the introduction of multi-party elections is insufficient, as they are susceptible to manipulation by incumbents, and it takes time to build other essential institutions, such as an independent judiciary and the media. As Hegre *et al.* (2001) show over a long historical period, the risk of civil war is greatest in poorly-designed democracies that retain authoritarian characteristics. Moreover, the quality of fiscal institutions may be so low as to be unable to deliver the necessary transfers even if the society is resolved to build a new social contract. This is a problem in Nigeria for example, in rebuilding from military dictatorship.

The chances of peace enduring are better if economic growth is rapidly restored, because then the contest is over a growing, rather than static or declining pie. The tax base rises with growth, thereby increasing the possibilities of redressing grievances through fiscal expenditure and transfers, provided that the requisite investments in fiscal institutions are made. In Afghanistan, for example, fiscal institutions are so degraded that there is not even a proper finance ministry to receive aid. Growth also tightens the labour market, thereby reducing the relative remuneration of joining a warlord or rebel group (Collier and Hoeffler, 2001: 6).

But growth by itself is insufficient to guarantee lasting peace. Growth must be broad-based, lowering poverty, but it must also reduce inter-group inequalities, the horizontal inequalities that are at the root of conflict. Addressing these requires deep systemic institutional change and design. Additionally, there are three serious impediments to restoring growth in post-conflict economies: (a) the physical destruction of infrastructure; (b) the emergence of new transactions costs in production; and, (c) a rise in the discount rate deterring investment. The last two are particularly problematic. We now turn to some of these issues.

Growth with redistribution

The initial years of peace may offer considerable scope for growth, as the economy re-achieves its pre-war production frontier. Thus annual real GDP growth rates of 8–10 per cent, such as achieved after the end of Lebanon's civil war, are not uncommon in post-conflict countries. But just as economic growth in general can distribute broadly or narrowly its benefits across society – depending upon the initial distribution of assets and skills – so too does reconstruction-led growth. Pre-war asset and skill distributions may have been highly unequal (with the resulting grievances contributing to conflict), and can worsen dramatically during wartime. The already poor often lose the few assets they have, and looting adds to the number of poor (Nafziger *et al*. 2000). In contrast, warlords and their followers accumulate assets. Therefore, while the early years of peace may see quite rapid GDP growth, it can be very narrow in its benefits – unless policies are put in place to restore the productive assets and human capital of the poor. The immediate post-conflict situation may offer a golden opportunity for pro-poor asset redistribution as well

(something that cannot be done easily during peace), although this can be impeded when rich 'winners' from war block the necessary measures.

Transactions costs in production

The 'post-conflict' economy will be highly distorted, and this can impose an unfortunate path-dependence on reconstruction and growth. One source of distortion is the sharp increase in transactions costs resulting from war, including the destruction of transport, the planting of land mines, and institutional collapse, that drive a wedge between producer and consumer prices. Typically, production (especially agriculture) is more vulnerable, leading to a sharper increase in its transactions costs compared to other sectors such as urban-based trade and services (the case in Mozambique's sixteen-year civil war, for instance). This effect is additional to the increase in production costs resulting from the direct destruction of institutions and infrastructure which usually disproportionately affects production, especially agriculture. Services and trade such as profitable smuggling and other rent-seeking heavily dominate wartime economies and, because of the relative price effect, the collapse in production usually far exceeds that due to destruction alone. Smallholder agriculture – the basic livelihood of Africa's poor – is hit especially hard. Angola's agricultural output, for example, is now less than five per cent of its pre-war level.

If a peace deal is signed then transactions costs start to fall. Accordingly, agricultural output recovers as relative prices shift back in its favour, and as infrastructure is rebuilt. Although this reconstruction may be characterised by a more than proportionate rise in sectors such as agriculture, recovery's benefits may still be too narrowly distributed due to the loss of assets and skills among the poor (see de Sousa, 1999 on Mozambique). Moreover, the economy's pre-war structure may have been far from optimal for growth, poverty reduction, or peace due to policy distortions. Hence, significant economic reform must take place alongside reconstruction if broad, rather than narrow recovery is to take place. This is not to say that reform itself is straightforward. There is widespread agreement on the need for fiscal reform to raise public spending on pro-poor basic health care and primary education, but other reforms are hotly debated. For example privatisation has had mixed success, especially in conflict countries when the powerful, including war-criminals, have taken over state assets.

High discount rates

Note that if peace is uneasy then there is an additional constraint on rebuilding the social contract via broad-based reconstruction and growth. The rise in the private rate of discount acts as a greater deterrent to investing in production sectors. These require more fixed capital than trade and services (which may only need working capital), and profits from trade can be more quickly realised than production whose profits may only start to flow after several periods (particularly in agriculture, where tree crops have lengthy gestation periods). Hence, any rise in the private discount

rate will tend to have a more depressing effect on the expected profitability of investments in production as compared to investments in services. Restarting activity in pro-poor production activities such as agriculture can therefore be especially hard after war when the peace is uneasy, for while transactions costs can be reduced and infrastructure rebuilt, an uneasy peace keeps private discount rates high. Consequently the reconstruction-growth path may create insufficient income and employment, its narrowness leading to frustration amongst large sections of the excluded population. Hence, the effectiveness of donor support to reconstruction will be reduced if parallel political initiatives (both domestic and international) are not undertaken to secure peace, and lower investor uncertainty.

Conclusions

The causes of civil wars are complex. Colonial history and external intervention are crucial, but economic factors are certainly at work. It has to be borne in mind that both individual and group-level violent expropriation is often a viable alternative to peaceful economic production. Since civil war ultimately lowers growth and economic prosperity, conflict prevention and management is central to the strategy of economic development. This chapter has emphasised the role of grievance, relative deprivation and the contest over natural resource wealth in triggering civil war. The high value of certain categories of resource-rents that are easily capturable, and the relatively low private cost of engaging in their capture, motivate many civil wars in Africa and elsewhere. The credibility and reputation of parties to conflict matters in devising peace. Belligerents may eventually be keen to make a peace deal, but mutual suspicion and the difficulty of signalling commitment may impede progress for a considerable time. Finally, reconstruction is not simply a matter of rebuilding shattered infrastructure. That is far easier than the daunting task of recreating a workable social contract. For this requires both close attention to institutional reform and constitutional re-design, deep interventions that reduce grievances. If the social contract remains weak, then investor uncertainty will remain high, the economy's structure will remain distorted, and grievances will continue to fester. It has to be remembered that the breakdown of the social contract is a reflection of state failure, a problem that is more acute in nations whose formation is artificial, and whose institutions of conflict management and contract enforcement are degraded. Ultimately lasting peace is only possible when the intergroup inequalities that are the foremost sources of grievances are addressed by lowering inequality. Thus, reconstruction strategies must not only increase broad-based growth and reduce poverty, but also address relative income gaps.

Acknowledgement

We are grateful to Dorothy Rosenberg and Bill Yates for comments on previous drafts.

Notes

1 These factors are not limited to the developing world. In Northern Ireland sectarianism involved job discrimination. Closely-knit and relatively deprived communities provided the soldiery for the paramilitary groups. Later the greed–grievance nexus produced more grievance and greed. Contributions from the diaspora in North America, as well as criminal activity, now play an important part in financing Northern Ireland's sectarian extremism and violence.

2 The measurement of conflict and its occurrence is not straightforward (see Ayres, 2000, Wallensteen and Sollenberg, 2000). The Carnegie Commission on Preventing Deadly Conflict (1997: 12) lists the following 39 countries as having had conflicts which led to at least 1000 deaths in any one year in the 1990s: Afghanistan, Algeria, Angola, Azerbaijan, Bosnia and Herzegovina, Burundi, Cambodia, Chad, Chechnya, Republic of Congo, Democratic Republic of Congo (Zaire), Colombia, Croatia, East Timor, El Salvador, Eritrea (secession from Ethiopia in 1993), Georgia, Ghana, Guatemala, Haiti, Iraq, Kashmir (India v. Pakistan), Lebanon, Liberia, Moldova, Nigeria, Persian Gulf War, Peru, Rwanda, Sierra Leone, Sindh, Somalia, South Africa, southeast Turkey (Kurdistan), Sri Lanka (Tamil uprising), Sudan, Tajikistan, Uganda, Yemen. To this list we add: Burma, Irian Jaya (Indonesia), Kosovo, and the 1998–2000 Eritrea–Ethiopia war to produce our total of 43 conflicts.

3 Kaldor (1998: 120) develops this theme. She writes: 'Since the power of the warring parties depends largely on fear and/or self-interest and not on consent, they need an insecure environment to sustain themselves both politically and economically. Politically, identity is based on fear and hatred of the other; economically, revenues depend on outside assistance for the war effort and on various forms of asset transfer based on loot and extortion or on price distortions resulting from restrictions on freedom of movement. In peacetime, these sources of sustenance are eroded.'

4 In her sample of 29 cases of full-scale civil war between 1940 and 1992, peace lasted more than five years in only eight out of 17 cases in which the belligerents had embarked on serious negotiations to end hostilities (Walter, 2001).

References

Addison, T. (2000a) *From Conflict to Reconstruction in Africa*, Discussion Paper 2001/16, Helsinki: UNU/WIDER.

Addison, T. (2000b) 'Aid and Conflict', in Finn Tarp (ed.) *Foreign Aid and Development*. London: Routledge.

Addison T., P Le Billon and S.M. Murshed (2000) 'On the economic motivation for conflict in Africa', paper given at the Annual World Bank Conference on Development Economics, Paris, 26 June 2000.

Addison, T. and S.M. Murshed (2001a) *Credibility and Reputation in Peace Making*, Discussion Paper 2001/45, Helsinki: UNU/WIDER.

Addison, T. and S.M. Murshed (2001b) *The Fiscal Dimensions of Conflict and Reconstruction*, Discussion Paper 2001/49, Helsinki: UNU/WIDER.

Arunatilake, N., S. Jayasuriya, and S. Kelegama (2000) *The Economic Costs of the War in Sri Lanka*, Research Studies: Macroeconomic Policy and Planning Series 13, Colombo: Institute of Policy Studies.

Auty, R.M. (1998), 'Resource abundance and economic development, improving the performance of resource-rich countries', *Research for Action*, 44, Helsinki: UNU/WIDER.

Ayres, R.W. (2000) 'A world flying apart? Violent nationalist conflict and the end of the cold war', *Journal of Peace Research*, 37, 105–17.

Azam, J.-P. (2001) 'The redistributive state and conflicts in Africa', *Journal of Peace Research*, 38, 429–44.

Azam, J.-P. and A. Hoeffler (2001) *Violence Against Civilians in Civil War: Looting or Terror*, Discussion Paper 2001/46, Helsinki: UNU/WIDER.
Azam, J.-P. and A. Mesnard (2001) 'Civil war and the social contract', paper presented at the Royal Economic Society Annual Conference, Durham, 9–11 April.
Balencie, J.-M. and A. de La Grange (1999) *Mondes Rebelles: Guerres Civiles et Violences Politiques*, Paris: Michalon.
Berdal, M. and D. Malone (eds) (2000) *Greed and Grievance: Economic Agendas in Civil Wars*, Lynne Rienner, Boulder CO.
Carnegie Commission on Preventing Deadly Conflict (1997) *Preventing Deadly Conflict*, New York: Carnegie Corporation.
Collier, P. and A. Hoeffler (1999) 'Justice-seeking and loot-seeking in a civil war', World Bank, unpublished manuscript.
Collier, P. and A. Hoeffler (2001) 'Aid policy and peace: reducing the risks of civil conflict', paper presented at the Royal Economic Society Annual Conference, Durham, 9–11 April.
Crisp, J. (1998) *The 'Post-Conflict' Concept: Some Critical Observations*, Geneva: Research and Development Policy Unit, United Nations High Commission for Refugees.
de Sousa, C. (1999) *Rebuilding Rural Livelihoods and Social Capital: the Mozambique Experience*, Working Paper 171, Helsinki: UNU/WIDER.
de Soysa, I. (2000) 'The resource curse: are wars driven by rapacity or paucity?', in M. Berdal and D. Malone (eds) *Greed and Grievance: Economic Agendas in Civil Wars*, Boulder CO: Lynne Rienner.
Dixit, A.K. (1996) *The Making of Economic Policy: A Transaction-Cost Politics Perspective*, London: MIT Press.
Dixit, A.K. (1999) 'Some lessons from transaction-cost politics for less-developed countries', unpublished manuscript, Princeton University.
Edgeworth, F.Y. (1881) *Mathematical Psychics*, London: C. Kegan Paul.
Ellman, M. and L. Wantchekon (2000) 'Electoral competition under the threat of political unrest', *Quarterly Journal of Economics*, CXV (2): 499–531.
Findlay, R. (1996), 'Towards a model of territorial expansion and the limits of empire', in M.R Garfinkel and S. Skaperdas (eds) *The Political Economy of Conflict and Appropriation*, Cambridge: Cambridge University Press.
Garfinkel, M.R. and S. Skaperdas (2000) *Contract or War? On the Consequences of a Broader View of Self-Interest in Economics*, Irvine CA: University of California-Irvine.
Grossman, H.I. (1991) 'A general equilibrium model of insurrections', *American Economic Review*, 81, 912–21.
Gurr, T. (2000) 'Ethnic warfare on the wane', *Foreign Affairs*, 79 (3): 52–64.
Hegre, H., T. Ellingsen, S. Gates and N.-P. Gleditsch (2001) 'Towards a democratic civil peace? Democracy, civil change, and civil war 1816–1992', *American Political Science Review*, 95, 17–33.
Hirshleifer, J. 1995. 'Anarchy and its breakdown', *Journal of Political Economy* 103, 26–52.
IMF (1999) *World Economic Outlook*, Washington DC: IMF.
Kaldor, M. (1998) *New and Old Wars. Organized Violence in a Global Era*, Cambridge: Polity Press.
Keen, D. (1998) *The Economic Functions of Violence in Civil Wars*, Adelphi Paper 320, London, Oxford: IISS and Oxford University Press.
Klugman, J. (1999) *Social and Economic Policies to Prevent Complex Humanitarian Emergencies: Lessons from Experience*, Policy Brief 2, Helsinki: UNU/WIDER.
Kovsted, J. and F. Tarp (1999) *Guinea-Bissau: War, Reconstruction and Reform*, Working Paper 168, Helsinki: UNU/WIDER.
Kumar, K. (1998) 'After the elections: consequences for democratization', in Krishna Kumar (ed.) *Postconflict Elections, Democratization, and International Assistance*, Boulder, CO: Lynne Rienner.
Lashmar, P. (2000) 'Every diamond has human blood on it', *The Independent* (London), 24 May: 14.

Le Billon, P. (1999), *A Land Cursed by its Wealth? Angola's War Economy 1975–99*, Research in Progress 23, Helsinki: UNU/WIDER.

Mehlum, H., K.O. Moene and R. Torvik (2001) *The Market for Extortions*, Discussion Paper 2001/26, Helsinki: UNU/WIDER.

Murshed, S.M. and M. Perälä (2001) 'Does the type of natural resource endowment influence growth?', unpublished manuscript, Helsinki, UNU/WIDER.

Murshed, S.M. and S. Sen (1995) 'Aid conditionality and military expenditure reduction in developing countries: models of asymmetric information', *Economic Journal*, 105: 498–509.

Nafziger, E.W., F. Stewart and R. Väyrynen (eds) (2000) *War, Hunger, and Displacement: the Origins of Humanitarian Emergencies – Volume 1: Analysis*, Oxford: Oxford University Press for Queen Elizabeth House, International Development Centre University of Oxford and UNU/WIDER.

Nayyar, D. (ed.) (2002) *Governing Globalization: Issues and Institutions*, Oxford: Oxford University Press/WIDER.

Olson, M. (1996) 'Big bills left on the sidewalk: why some nations are rich, and others poor', *Journal of Economic Perspectives*, 10 (2): 3–24.

Paige, J.M. (1997) *Coffee and Power: Revolution and the Rise of Democracy in Central America*, Cambridge MA: Harvard University Press.

Polachek, S., J. Robst and Y. Chang (1999) 'Liberalism and interdependence: extending the trade-conflict model', *Journal of Peace Research*, 36, 405–42.

Reno, W. (1998) *Warlord Politics and African States*, Boulder CO: Lynne Rienner.

Rodrik, D. (1998) *Where Did all the Growth Go? External Shocks, Social Conflict, and Growth Collapses*, Cambridge MA: Harvard University, John F. Kennedy School of Government.

Skaperdas, S. (1992) 'Cooperation, conflict and power in the absence of property rights', *American Economic Review*, 82, 720–39.

Skaperdas, S. (2002) 'Warlord competition', *Journal of Peace Research*, 39, forthcoming.

Stewart, F. (2000) 'Crisis prevention: tackling horizontal inequalities', *Oxford Development Studies*, 28, 245–2.

Wallensteen, P. and M. Sollenberg (2000) 'Armed conflict 1989–1999', *Journal of Peace Research*, 37, 635–49.

Walter, B.F. (2001) *Committing to Peace: The Successful Settlement of Civil Wars*. Princeton: Princeton University Press.

Wedel, J. (2001) *Clans, Cliques and Captured States: Rethinking Transition in Central and Eastern Europe and the Former Soviet Union*, Discussion Paper 2001/58, Helsinki: UNU/WIDER.

Zartman, I.W. (ed.) (1995) *Collapsed States: The Disintegration and Restoration of Legitimate Authority*, Boulder, CO: Lynne Reinner.

4 Russia

The political economy of transition in a mineral economy

Richard M. Auty

Russia is a mineral economy. It has 5 per cent of world oil reserves and one-third of the natural gas reserves (*Economist*, 1999). However, the standard definition of a mineral economy is based not on the scale of the resource endowment but upon the *relative* importance of minerals to the economy. It states that a country is a mineral economy if it derives at least 10 per cent of its GDP and 40 per cent of its exports from minerals (Nankani, 1979). Although the Russian mineral sector's share of GDP almost halved during the years 1992–97 to 16 per cent of GDP, it remained within the definition threshold. As for exports, minerals became dominant because exports of other primary products and manufactured goods fell more dramatically. Fuel exports alone accounted for 42.3 per cent of total exports 1996–98 (World Bank, 2000). Finally, the contribution of minerals to total taxes rose despite the declining share of natural resources taxation in GDP. This is because the overall tax base shrank even faster.

Natural resource abundance can facilitate economic development if the rents are used to boost the levels of investment and taxation, and the foreign exchange from resource exports enhances import capacity. However, evidence from the developing market economies (Sachs and Warner, 1995; Auty, 2001a) and more recently from the transition economies (Auty 2001b) indicates that natural resource rents can repress economic development. For example, World Bank data on natural resource rents, available for the year 1994 only, show an inverse relationship between natural resource rents and per capita GDP growth during the years 1985–97 (Table 4.1). Among four categories of resource-abundant country, the mineral economies had the highest rents in relation to GDP but the slowest per capita GDP growth.

Gaddy and Ickes (1998), and Markandya and Averchenkova (2001) analyse how mineral rents affect Russian economic reform, but they underestimate the impact because they do not conceptualise Russia as a mineral economy. This chapter fills the gap by applying a model of mineral-driven economic growth to analyse how the mining sector has impacted Russia's transition to a market economy. Russia is one of six mineral economies in transition; the other five (Mongolia, Azerbaijan, Kazakstan, Turkmenistan and Uzbekistan) are all much

Table 4.1 Share of rent in GDP 1994 and per capita GDP growth, six natural resource endowment categories

Resource endowment	*Pasture and cropland rent (% GDP)*	*Mineral rent (% GDP)*	*Total rent (% GDP)*	*Per capita GDP growth 1985—97 (%/yr)*
Resource poor				
Large	7.34	3.22	10.56	4.7
Small	5.41	4.45	9.86	2.4
Resource rich				
Large	5.83	6.86	12.65	1.9
Small, non-mineral	12.89	2.53	15.42	0.9
Small, hard mineral	9.62	7.89	17.51	–0.4
Small, oil exporter	2.18	19.04	21.22	–0.7

Source: Derived from World Bank (2000).

Notes:
Resource-poor = 1970 cropland/head < 0.3 hectares
Large = 1970 GDP > $7 billion

Table 4.2 Classification of transition economies by natural resource endowment

Category	*Mean GDP ($ Billion)*	*Cropland per capita (hectares)*	*Mineral exports (% GDP)*	*GDP 1997 as % 1989*	*Countries*
Large resource poor	929.0	0.08	6	218	China
Small resource poor	16.5	0.21	16	82	Armenia, Croatia, Czech Republic, Georgia, Kyrgyz Republic, North Korea, Slovak Republic, Slovenia, Tajikistan, Vietnam
Small resource rich	32.5	0.53	11	68	Belarus, Bulgaria, Estonia, Hungary, Latvia, Lithuania, Moldova, Poland, Romania, Ukraine
Mineral economies	78.4	0.73	57	65	Azerbaijan, Kazakstan, Mongolia, Russia, Turkmenistan, Uzbekistan

Source: World Bank (1999), and WRI (1998) for cropland.

smaller economies (Table 4.2). The relatively large size and apparently diversified structure of the Russian economy might be expected to mute the impact of mining on its development, but this chapter finds this is not the case.

The chapter begins by deploying resource-driven growth models to generate hypotheses about the Russian transition. More specifically, the second section examines the impact of the resource rents on the speed of reform. The third section measures the Dutch disease effects and their impact on economic restructuring.

The fourth section analyses the impact of the natural resource rents on rent-seeking behaviour and corruption. The fifth section examines the growth trajectory. Finally, the last section draws the conclusions and summarises the policy implications.

Mineral-driven models of economic growth and governance

A UNU/WIDER study (Auty 2001a) develops two basic models of resource-driven growth. The first is the competitive industrialisation model, which is most strongly associated with the resource-poor countries. The second is the staple trap model that has been associated with the resource-abundant countries since the 1960s. The competitive industrialisation model provides a useful counterfactual to the staple trap model.

The competitive industrialisation model summarises the development trajectory of the resource-poor countries under a developmental political state. The differentiating features of structural change in resource-poor countries are, first, an early start on *competitive* industrialisation compared with the resource-rich countries, and, second, reform at a low per capita income towards a relatively open trade policy that triggers an expansion of labour-intensive manufacturing, much for export. This development trajectory promotes virtuous interlocking economic and social circles that sustain rapid economic growth that is also equitable.

The virtuous *economic* circle exhibits four key characteristics. First, early and rapid industrialisation accelerates urbanisation. This speeds the demographic cycle so that the ratio of dependants to workers falls, stimulating saving and investment (Bloom and Williamson, 1998). Second, the rate of investment quickly rises towards 25 per cent of GDP. Most importantly, the efficiency of investment is sustained so that the per capita GDP can double each decade. Third, the arrival at the labour market turning point (which occurred within a decade after trade policy reform in both South Korea and Taiwan) creates pressure for wage increases that require the economy to diversify into competitive heavy and chemical industry (HCI). This diversification reduces the vulnerability of the economy to external shocks. Fourth, the rapidly increasing complexity of the economy fosters a 'linear liberalisation' that further constrains rent-seeking behaviour.

The interlocking virtuous *social* circle has three main features. First, income equality is maintained because the tightening of the labour market removes the drag of surplus rural labour on the wages of the poor. In addition, the upgrading of workforce skills in HCI reduces the wage premium on higher skills and compresses the income range (Londono, 1996). Birdsall *et al*. (2001) and also Wood and Berge (1997) confirm the stronger incentive for skill acquisition provided by structural change in resource-poor economies compared with resource-abundant ones. Second, the resulting relatively equitable distribution of income together with earlier urbanisation and passage through the demographic cycle helps lower transaction costs by boosting social capital accumulation compared with resource-abundant countries (Woolcock *et al*., 2001). Third, these trends in human and social capital combine to push the political economy towards a democracy that tends to be consensual, rather than polarised so that policy coherence is sustained but the risk of policy capture by vested interests is reduced.

The staple trap model of economic growth ascribes the disappointing performance of the resource-rich countries to two fundamental departures from the competitive industrialisation model. First, the pattern of structural change entails a late start on competitive industrialisation. Second, the political economy of resource-rich countries tends to engender predatory political states. More specifically, resource-abundant countries rely on primary product exports for longer than resource-poor countries and this postpones diversification into competitive manufacturing with four adverse consequences. First, economic diversification must initially occur into other primary products and this may prove difficult for the smallest resource-abundant economies whose natural resource endowment is likely to be skewed towards one or two viable commodities (Auty, 2001a). Second, slower industrialisation retards urbanisation so that passage through the demographic cycle is delayed and the favourable high-saving phase of the dependency/worker cycle is postponed and the accumulation of produced and human capital is slower. Third, the manufacturing that does develop tends to be capital-intensive so that it does not absorb surplus rural labour and the resulting depressed wage of the poor raises income inequality and also social tensions.

Lastly, there is a misallocation of entrepreneurial talent because rent-seeking is more remunerative than productive activity. Murphy *et al*. (1991) note that in the face of weak property rights and extensive government regulation, the ablest individuals are drawn into public rent-seeking where returns are higher than in the private sector due to the predatory behaviour of the state. Elsewhere, Murphy *et al*. (1993) argue that innovative private sector activity is particularly at risk because numerous government permits are required to set up a new business, each of which creates opportunities to extract rent. There may be increasing returns to scale in public sector rent-seeking so that it expands at the cost of economic growth. Rent-seeking also creates opportunities for corruption and Mauro (1995) provides cross-country evidence that such corruption represses economic growth. Taking a corruption index based on country risk studies that ranged from 0 (most corrupt) to 10 (least corrupt), Mauro (1995) reports that an improvement of two points on the scale increases investment by four per cent and the annual per capita GDP growth rate by 0.5 per cent. Mauro (1995) also finds that corruption is more prevalent in resource-abundant countries than in resource-poor ones. He attributes this to the fact that the natural resource rents and the sales of the natural capital assets that generate them are subject to state regulation that creates opportunities for political abuse. Leite and Weidmann (1999) also link corruption to natural resources. In addition, they show that corruption is higher if institutions are weak (a low score on the index of the rule of law), if political instability is high and if the trade policy is closed.[1]

Turning to the second basic divergence from the competitive industrialisation model, longer dependence upon primary products merely postpones competitive industrialisation if the political state is developmental. However, the political state in most resource-abundant economies tends to be predatory. This is because the presence of natural resource rents reduces the incentive for the governments of resource-rich countries to provide public goods to lift output in order to advance their interests. Assuming there are two sources of income available to governments,

resource rents and returns to investment, then low rent and high-rent countries experience contrasting outcomes. The high-rent government will find it easier to satisfy its financial needs by capturing the rents than by investing to generate wealth so that the latter will be neglected in favour of the former. Effort will be diverted into the political process by which rent is extracted, and away from measures to raise productivity, like improving institutions. Indeed, effective institutions may be regarded as an impediment to rent-seeking behaviour because they eliminate rents by increasing government accountability and promoting competition. In contrast, governments of resource-poor countries have an interest in improving institutions in order to increase their financial resources. The leadership maximises its remuneration by maximising national output.[2]

Predatory governments in resource-rich countries deploy the resource rents at the expense of a coherent economic policy and long-run social welfare. Competitive industrialisation is thereby indefinitely postponed rather than delayed. This is because impatience with the rate of employment creation under resource-driven growth leads to forced industrialisation by trade policy closure to protect infant industry. Dutch disease[3] effects from resource booms are a secondary cause of trade policy closure (Murshed, 2001; Sachs, 1999). The policy of infant industry protection has three principal flaws. First, the policy creates discretionary rents that governments deploy with minimal transparency so that they degenerate into a corrupt rent-dispensing mechanism that distorts the economy in an *ad hoc* and cumulative fashion. Second, infant industry is capital-intensive and creates few jobs so that governments feel obliged to provide additional employment that is often non-productive. This relaxes market discipline and expands still further the protected sector. Third, the policy tends to be captured by those who benefit from it so that the protected industry is slow to mature and imposes increasingly onerous demands on the primary sector for transfers and foreign exchange. The transfers outstrip the rents and depress incentives within the primary sector (Krueger and Tuncer, 1982). The transfer of rents into a burgeoning non-tradeable sector that includes slow-maturing infant industry and non-productive public sector employment causes a steep decline in investment efficiency, so that the level of investment flattens and may fall.

The staple trap model is also associated with worsening income inequality and the slow accumulation of social capital. This is because the labour market fails to expand fast enough to absorb surplus rural labour so that the remuneration of the poorest is depressed while a unionised labour aristocracy boosts the wages of the favoured, but relatively few, workers in the protected capital-intensive sector. Meanwhile, rent-seeking further skews the distribution of income and assets in favour of a politically powerful minority. Social tensions arise, exacerbated by decelerating economic growth, so that they intensify as the economy deteriorates. In this way the resource-driven economy becomes locked into a staple trap of increasing dependence on commodities with declining competitiveness so that it is vulnerable to external shocks and a growth collapse from which recovery is protracted.

The adverse effects of resource abundance are heightened with ‘point’ natural resources, like minerals, compared to resources with diffuse socio-economic

linkages like peasant crops within a developmental political state (Baldwin, 1956; Bevan *et al*. 1987). The capital-intensive production function of mining concentrates revenues on one or two large firms, a small labour aristocracy of mineworkers and the government. Taxation therefore dominates domestic socio-economic linkages, enhancing the role of the government and, consequently, raising the probability of policy failure. Extreme resource abundance lengthens the period of economic distortion (foreign aid can have a similar effect).

Consistent with the staple trap model, central planning forced industrialisation in Russia in a manner that eventually encountered diminishing returns. When the system collapsed, the social entitlements that it had supported collapsed also. This means that application of the resource-driven growth model to the transition economies requires adjustments to reflect the singular institutional legacy of central planning and also the inherited gap between the demand for public services under central planning and the government revenue subsequently available to meet that demand.

A useful proxy for the quality of institutions is provided by De Melo *et al*. (1997) in terms of the degree of suppression of market transactions and the institutions that facilitate these transactions. The distortion is conditioned by the length of time that an economy has been exposed to central planning. The index ranges from -0.70 for the countries of Central and Eastern Europe and the Baltic Republics (CEE), which are the least distorted to an index of 1.12 for the mineral economies of the Caspian region, which are the most distorted. The index for Russia is 0.32, implying significantly greater distortion than the CEE countries but slightly less distortion than the East Asian countries (0.52). However, Russia and the other northern Commonwealth of Independent States (CIS) countries face the most daunting combination of long exposure to central planning and a large endowment of obsolete physical capital.

Turning to the public service funding gap, the gap originates from the fact that central planners were able to repress consumer demand in order to divert a larger fraction of output to public goods provision than a market economy at a comparable stage of development. The collapse of central planning sharply compressed the productive capacity of the economy. The faster CEE reformers lost productive capacity equivalent to one-fifth of GDP and the associated capacity to provide public goods. Elsewhere, the more distorted CIS economies lost on average almost half their output, and in some cases over three-fifths of GDP, while the share of government revenue in GDP also shrank, thereby exerting massive compression on the provision of public goods (EBRD, 1999: 58). The collapse in Russia was two-fifths of GDP, twice that in the CEE countries.

The mineral-driven model of economic development suggests (Auty, 2001b) that, compared with resource-poor countries, mineral-rich countries in transition are more likely to experience:

- slower accumulation of institutional capital and slower implementation of reform;
- faster rebound of the real exchange rate and hence slower economic restructuring;

- higher levels of rent-seeking and corruption due to weaker sanctions against anti-social governance;
- steeper falls in GDP tax revenues and lagged resumption of economic growth, especially in the presence of high post-communist compression of GDP.

The next four sections test the basic hypotheses generated from the models against Russian experience.

Did natural resource rents retard institutional and economic reform?

The basic pattern of Russian reform is one of an early start that subsequently stalled. Table 4.3 indicates the extent to which Russia lags the CEE countries, especially in financial reform, trade reform and the restructuring of government and enterprises. Only in privatisation did the country almost close the gap, but with far from positive consequences as is shown below. Moreover, it is interesting to note that the slow pace of Russian reform is on a par with that of the lower-income resource-poor Caspian Basin countries. The consequences of lagged reform for institutional change are summarised in Table 4.4: overall, Russia lags the CEE countries by a substantial amount and, in terms of institutional quality has a similar index to that of the low-income gradual reformers of East Asia.

An important reason for the relatively high Russian index for privatisation (Table 4.3) is that the reformers used the early sale of state-owned natural resource monopolies to boost post-privatisation tax revenues instead of generating income from the sale of keenly-priced assets. In line with this strategy the reformers set relatively high levels of taxes but this merely encouraged ingenious forms of evasion (Goldman 1999). For example, in 1992 the government had expected $10 billion from oil export taxes alone, but the actual outcome was minimal due to evasion (Boone and Fedorov, 1997: 183).

Poland pursued a more successful strategy with a more gradual approach to tax reform during 1990–92 that elicited a firm commitment from managers to restructure state-owned enterprises (SOEs) over a two to three year period. The Polish strategy involved sequenced closure of faulty policy instruments, starting with the elimination of explicit government subsidies, followed by the abolition of soft credits and then of inter-enterprise arrears, and ending with the elimination of tax arrears. Unlike Poland, Russian reformers failed to generate sufficient revenue to narrow the fiscal gap and stabilise prices, a key pre-condition for sustained reform. Shleifer and Treisman (2000) attribute the flawed reform strategy to the inability of the reformers to build a political coalition in favour of stabilisation. The beneficiaries of inflation were too powerful and included the central bank and many newly-formed commercial banks, along with state-subsidised enterprises, farms and organisations (notably the military).

Attempts to stabilise the economy failed until 1995 when the natural resource rents were deployed once again, this time to co-opt sufficient opponents of reform to support stabilisation. The energy companies lie at the heart of the deal by which

Table 4.3 EBRD transition indicators 1999

Country/Group	*Privatisation*	*Govt and firm restructure*	*Price liberalisation*	*Trade liberalisation*	*Competition policy*	*Financial reform*	*Overall index*
Lower income							
East Asia	2.0	2.1	2.9	2.5	1.8	1.5	2.1
Resource-poor Central Asia	3.2	1.9	3.0	3.7	1.8	1.7	2.5
Resource-rich Central Asia	2.6	1.9	2.5	2.1	1.5	1.7	2.1
Northern CIS plus SEE	2.9	1.8	2.7	3.2	2.1	2.0	2.5
Russia	3.7	1.7	2.7	2.3	2.3	1.7	2.5
Mid income							
Central and Eastern Europe	3.9	2.9	3.1	4.3	2.7	3.0	3.3

Source: IMF (2000), 134.

Note: The privatisation and financial reform indices are collapsed from two readings into one, but the aggregate index is derived from equally weighting all the original eight readings.

Table 4.4 Institutional quality index, transition economies 1997–98

Country/Group	*Voice and accountability*	*Rule of law*	*Effective governance*	*Political stability*	*Regulation burden*	*Graft*	*Overall index*
Lower income							
East Asia	–11.7	–4.8	–1.4	5.7	–6.0	–3.1	–5.8
Resource poor Central Asia	–5.2	–6.1	–9.0	–5.8	–10.3	–9.1	–7.4
Resource-rich Central Asia	–11.1	–7.5	–10.5	–5.4	–10.5	–10.3	–8.8
Northern CIS and SEE	–1.1	–3.0	–6.6	–1.8	–3.3	–6.0	–3.6
Russia	–3.1	–7.2	–5.9	–6.9	–3.0	–6.2	–5.4
Mid Income							
Central and Eastern Europe	8.8	4.2	3.4	7.4	4.8	2.7	5.0

Source: IMF (2000), 136.

Table 4.5 Non-payments, Russia 1994–98 (% GDP)

	1994	*1995*	*1996*	*1997*	*1998*
Total arrears	14.8	15.1	23.4	19.2	39.3
To suppliers	9.2	7.7	11.2	13.3	17.7
To budget	3.2	4.7	9.2	12.2	16.4
Wage arrears	0.8	0.9	1.6	1.5	2.9
Net Cash Sales/ Sales (%)	17	22	35	41	50

Source: Pinto *et al*. (1999): 2.

opponents were co-opted to support stabilisation. The Russian government won over the banks by raising loans from them and offering state-owned assets as collateral. This effectively allowed the government to respond to depressed tax revenues by bringing tax revenues forward. When the state could not service the loans, the banks were able to auction off the energy assets so that they and their oil industry agents (mostly leading civil servants from Soviet times) acquired such assets easily and cheaply.

The SOEs were co-opted to stabilisation by the government requiring energy firms to supply customers like the SOEs, despite payments arrears, in exchange for tolerating non-payment of taxes by the energy companies. Pinto *et al*. (1999: 18–19) estimate that the energy sector provided subsidies to the economy of around 4 per cent of GDP annually 1993–97.[4] Gazprom alone, in which the Russian government held 37.5 per cent of the shares, provided an implicit subsidy to domestic (and CIS) consumers that averaged 1.6 per cent of GDP 1993–7. The gas and power companies by law could not disconnect debtor customers so they had to resort to alternative means to continue trading while absorbing arrears. For example, Gazprom received generous tax treatment notably on its lucrative foreign sales. Gazprom drew two-thirds of its revenues from exports to the market economies although those markets accounted for only one-quarter of its total output. It drew barely 12 per cent of its revenue from the 60 per cent of its sales made on the domestic market. The power utilities functioned similarly. Table 4.5 traces the growing share of tax arrears in Russian GDP: one-third of the utilities' payables was owed to Gazprom. The utilities provided subsidies estimated at 2.3 per cent of GDP annually 1993–7, but had less scope than Gazprom to recoup the subsidy that they conferred on other sectors.

In this way the energy companies' rents were used to subsidise obsolete SOEs by absorbing either barter payments from creditors that grossly over-value the goods so sold, or non-payments. However, the energy companies passed the cost of subsidies to the fiscal accounts in the form of non-payment and offsets (mutual cancellation of tax liability and budgetary arrears). Consequently, even as explicit government subsidies to the SOEs declined, implicit subsidies in the form of cheap energy and tolerance of non-payment of taxes were stepped up (Pinto *et al*. 1999). Table 4.5 shows that arrears are estimated to have almost tripled to two-fifths of GDP over the years 1994–98.

This system of energy subsidies jeopardised economic growth in two ways. First, it reduced incentives for enterprise restructuring. Second, it postponed effective stabilisation and tax reform. This led to a shortage of cash revenues that raised interest rates and thereby squeezed profit margins in a vicious circle that created government debt that was unserviceable. High interest rates also caused a sharp real appreciation of the exchange rate and retarded structural change, as the next section shows.

Did Dutch disease effects retard Russian restructuring?

Mineral economies are likely to experience an amplified rebound of the post-transition appreciation of the real exchange rate (Rosenberg and Saavalainen, 1998). This is caused by the inflow of foreign exchange for investment to revitalise the mining sector. A stronger real exchange rate weakens the competitiveness of the non-mining tradeables sector, mainly agriculture and labour-intensive manufacturing, and so discourages investment in those sub-sectors. Yet, although Kazakstan and Azerbaijan conform to this projection, Russia failed to attract an inflow of foreign capital on a similar scale, relative to its GDP. The per capita inflow of foreign capital into Russia over 1992–98 was $98, barely one-fifth that of Kazakstan. This partly reflects the nationalistic stance of the Russian government towards mining, but also its failure to stabilise the economy.

Nevertheless, the Russian real exchange rate did appreciate strongly after 1994 due to the very high interest rates required to stabilise prices in the face of the mounting government and corporate debt described in the previous section. The real appreciation of the exchange rate was therefore a direct consequence of the government's deployment of the energy rents. Gurgen *et al*. (1999: 44) estimate the real exchange rate of Russia doubled between 1994 and 1998, compared with a rebound in Kazakstan of 2.5 and a strengthening of 25 per cent for Kyrgyzstan and virtually no change for Uzbekistan. Consistent with this, the IMF (1999: 100) estimates that the Russian dollar wage recovered to $200 in 1998 compared with $150 in December 1995 and $100 in mid-1994, when it was broadly in line with that of other transition economies. Since neither labour productivity nor the dollar price of goods rose, the profitability of manufacturing was strongly squeezed.

The extent of the Dutch disease effect caused by the appreciation of the real exchange rate has been estimated by the EBRD (1997: 62). It uses data from forty-one market economies to establish a counter-factual of per capita income and structural change for the transition economies. The aggregate figures for the transition economies as a whole suggest that prior to reform the share of industry in total employment was 12.8 per cent higher than for market economies at a similar level of development, a size half as big again. Other sectors were correspondingly smaller, with the share of agriculture down by one-fifth (equal to 5.5 per cent of total employment), that of market-oriented services down by one-sixth (4.3 per cent of all employment) and that of other services down one-tenth (2.0 per cent of all employment). Paradoxically, a study of revealed comparative advantage by the EBRD (1999: 178–80) shows that few transition economies had a comparative

Table 4.6 Change in employment composition, by sector 1990–97 (% total)

Reform group	*Agriculture*	*Industry*	*Non-market services*	*Market services*
Central and East Europe	14 → 13	41 → 35	24 → 20	22 → 31
Southeast Europe	23 → 32	42 → 33	20 → 15	16 → 19
Central Asia and Caucasus	31 → 44	37 → 15	17 → 25	14 → 16
Northern CIS	21 → 26	41 → 28	21 → 24	16 → 21
Russia	*13 → 14*	*42 → 32*	*26 → 30*	*17 → 24*

Source: EBRD (2000).

advantage in the heavy industry and agriculture that central planning favoured. Rather, their comparative advantage lay in labour-intensive industry and natural resources. The revealed comparative advantage of Russia is strongly concentrated on natural resources, with only a very modest advantage in heavy industry.

Consistent with the pattern of revealed comparative advantage, Table 4.6 shows that the initial transition adjustment (1990–97) caused the share of industrial employment to contract substantially in all countries. However, only at relatively low-income levels did it tend to converge on the expected (market economy) level, and in some such cases it even fell below it. The distinctive pattern of structural change at high-income levels in the CEE countries reflects a production structure that is associated with wealthier market economies and suggests a continuing significant under-valuation of their per capita income. In contrast, the tradeable sectors of Russia and the two other oil-exporting countries (Azerbaijan and Kazakstan) contracted to levels below the expected share of employment (Table 4.7). This outcome is consistent with Dutch disease effects. Elsewhere, the resource-poor countries of the Caspian Basin, which also rely strongly on the primary sector, saw agriculture more than offset a sharp fall in industrial employment (Table 4.7, final line).

The energy subsidies did not offset lost competitiveness so the share of tradeables in Russian GDP was relatively low and, as noted, concentrated on minerals. Yet the mineral sector on which the economy increasingly relied had few incentives for long-term re-investment. Consistent with the mineral-driven staple trap model, this skewed and unstable export structure rendered the economy vulnerable to external shocks. The shock duly struck when prices of metals and hydrocarbons collapsed and the country's terms of trade fell by one-third between January 1997 and August 1998. Exports stagnated and the trade gap and current account each deteriorated by 7 per cent of GDP, causing the confidence of financial markets to collapse.

Pinto *et al*. (1999) attribute the 1998 financial crisis to failed fiscal reform, aggravated by contagion from the 1997 Asian financial crisis, a fragile banking system and the negative oil shock. However, this assessment under-estimates the critical role of natural resource rents. The sharp contraction in the oil rents in 1998 precipitated the financial crisis. Meanwhile, the rents had earlier institutionalised soft budgets and delayed tax reform and structural change so that economic

Table 4.7 Dutch disease effects 1990–97 (actual and expected employment share)

	Index	*1990*	*1997*	*Departure from norm in 1997*
Energy-rich				
Azerbaijan	Agriculture	0.31	0.29	–0.13
	Industry	0.25	0.14	–0.05
	Distortion	+0.03	–0.18	
Kazakstan	Agriculture	0.22	0.25	–0.03
	Industry	0.31	0.18	–0.05
	Distortion	+0.05	–0.08	
Turkmenistan	Agriculture	0.42	0.49	+0.06
	Industry	0.21	0.18	–0.02
	Distortion	+0.13	+0.04	
Russia	*Agriculture*	*0.13*	*0.14*	*–0.17*
	Industry	*0.42*	*0.32*	*+0.08*
	Distortion	*+0.03*	*–0.09*	
Resource-Poor				
Caspian Basin[1]	Agriculture	0.30	0.52	+0.12
	Industry	0.28	0.14	–0.06
	Distortion	+0.11	+0.11	

Source: EBRD (2000).

Note:
1 is for 1990–96

distortions were amplified. Moreover, the natural resource rents also fed rent-seeking behaviour and corroded institutional capital, as the next section shows.

Did minerals feed rent-seeking and corruption in Russia?

In fact, rent-seeking flourished in Russia before the transition to a market economy. Under the Law of Co-operatives that was established in 1988, the directors of SOEs were able to set up private firms within the SOEs and transfer profits in from the state sector, establishing a vested interest in rent-seeking. These rent-seeking opportunities were maximised when the reform stalled and the economy became stuck so that it combined arbitrary regulation with chaotic prices, interest rates and exchange rates. Åslund (1999) demonstrates this by establishing an inverse relationship between the speed of reform and the degree of rent-seeking. Conditions that nourished rent-seeking by a well-connected elite (the oligarchs) first emerged when Soviet finances collapsed in 1991 and the central government resorted to money emission while reform stalled. Reform briefly restarted 1994–95 and appeared to curb both rents and inflation before stalling once again and allowing rent-seeking opportunities to expand again (Table 4.8, final line). Åslund (1999) stresses that privatisation in itself did not create benefits for enterprise managers, rather it provided rent-seeking opportunities through the extraction of export rents, credit subsidies, import subsidies and other transfers from the state.

The energy sector did make a substantial contribution to rent-seeking behaviour. The domestic under-pricing of energy created sizeable opportunities for

Table 4.8 Key economic indicators, Russia 1992–98

	1992	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*
GDP ($ bn)	79	159	277	357	429	450	272
GDP growth (%/year)	–14.5	–8.7	–12.6	–4.2	–3.5	0.8	–4.6
Inflation (%/year)	2510	842	224	131	22	11	84
Unemployment (%)	4.9	5.5	7.5	8.2	9.3	9.0	11.8
Government revenue (% GDP)	39.2	36.2	34.2	31.0	31.0	32.6	32.0
Government expenditure	57.6	43.6	45.0	36.8	39.3	39.9	36.5
Fiscal deficit	–18.4	–7.3	–10.4	–5.7	–8.4	–7.4	–4.5
SOE subsidies (% GDP)	10.8	9.4	10.5	3.4	3.9	4.2	n.a.
Rents (% GDP)	81.0	35.0	15.0	8.0	11.0	15.0	5.0

Source: Åslund (1999), 58.

favoured groups. For example, in December 1991, the state controlled price of oil was $0.50 per tonne, less than 0.5 per cent of the world price. A this time, some 20,000 state firms were licensed to trade and they could purchase commodities at official prices, secure export licenses from officials and then sell goods at world prices. Significantly, efforts to liberalise prices in spring 1992 were effectively blocked by the hydrocarbon lobby on the grounds that Russian SOEs would be unable to cope with world prices. Consequently, domestic prices were still only 1 per cent of world prices in 1992 and averaged 8.3 per cent of world prices through 1993. Åslund (1999) reports that private firms owned by the directors of SOEs bought energy at low state prices from their SOEs and sold it at world prices, depositing the proceeds in offshore accounts. He estimates the total export rents at no less than $24 billion in 1992, a sum that amounted to 30 per cent of GDP, given the collapse of the rouble against the US dollar (Table 4.8).

There were three additional channels of rent-seeking.[5] The second channel was subsidised credits, which were issued in 1992 by the central bank with interest of 10–25 per cent when inflation was at 2,500 per cent. The value of the credits was 31.6 per cent of GDP of which 23 per cent of GDP went to the SOEs as a virtual gift. These benefits were less restricted than the export gains but they made many bankers wealthy. The third rent-seeking opportunity arose out of Russian credits to the associated republics in the rouble zone that cost the Russian Federation a further 22.5 per cent of GDP. The fourth transfer involved food subsidies that allowed a handful of merchants to import food at 1 per cent of domestic prices and sell it on the domestic market with little restriction at unregulated prices. The import subsidies were in the form of western export credits that were charged to Russian debt at 17.5 per cent of GDP in 1992. Rent-seeking experienced a resurgence when reform drifted during 1996–98, in contrast to a squeeze in 1994–95.

Western government fears about the risk of nuclear proliferation arising from the collapse of the Russian state may lie behind the achievement of an IMF standby

agreement in difficult circumstances. The agreement led to a large inflow of foreign direct investment and a much greater inflow of portfolio investment (matched by capital outflows of around $20 billion annually 1996–98). The capital inflows functioned like resource rents and led directly to relaxation of government tax-raising efforts while businessmen seized the assets of minority shareholders through a combination of transfer pricing, outright theft or by failure to service bond debt.

Gaddy and Ickes (1998) model Russia as a virtual economy in which rents and soft budget constraints support a large value-detracting manufacturing sector. This mechanism is at the heart of the staple trap model where a potentially dynamic primary sector (minerals) is exploited to sustain a cosseted manufacturing and government sector. However, Pinto *et al*. (1999) refine the virtual reality model because, if unmodified, it implies that industrial firms are the sole beneficiaries and that they benefit at the expense of energy firms, their workers and the government. They argue that this is not tenable after the initial round of expenditure within the virtual economy so that it should therefore collapse. The fact that it did not collapse is due to the ability of agents other than the SOEs to benefit from the system. For example, energy firms supported the system because it allowed them to diversify and accumulate assets by converting payment arrears into equity in the offending firms. Meanwhile, some firms like Gazprom could pass the net subsidy burden to the government in terms of tax arrears and unpaid funds to extra-budgetary accounts. Finally the government acquiesced in order to subsidise employment in failing SOEs.

Did minerals affect the transition development trajectory?

Table 4.9 shows that the CEE fast reformers experienced a shallower downturn in GDP compared to the CIS countries. More specifically, the CEE transition trajectory is one of a relatively modest decline in output, an early revival of growth and a steady recovery to seven-eighths of the pre-transition output within six years of reform. Significantly, within the Caspian Basin countries, the GDP rebound occurred later than in the CEE countries, but the rebound was stronger in the resource-poor countries and weaker in the mineral-rich countries. The latter experienced a more modest output collapse than the strife-ridden resource-poor countries but also a later and less robust recovery. The Russian trajectory resembles that of the Caspian mineral-rich countries.

In summary, the transition trajectory of the fast-reforming Central and East European countries traces a shallow U-shape compared with the reverse-J of the mineral-rich economies like Russia and the deep V-shape of the resource poor southern CIS countries. However, a 45 per cent real depreciation of the rouble in 1998 triggered a strong rebound in Russian manufacturing output. It also improved prospects in the oil sector. For example, the production costs of Lukoil, the largest company, fell to $3–4 per barrel stemming earlier financial losses (*Financial Times* 1999a, 1999b). This outcome is consistent with earlier Dutch disease effects and, unfortunately, without renewed institutional and economic reform; it is unlikely to be sustained.

Table 4.9 Reform and growth trajectories, Russia and regional groups

Country group	*Reform index*	*Cumulative GDP decline to low (%)*	*Year of lowest GDP*	*Cumulative GDP growth from low (%)*	*Annual GDP growth from low (%)*	*Ratio of GDP low to six-year transition GDP*
CEE + BR	3.3	28.9	1993	24.8	3.8	0.87
Northern CIS	2.5					
Russia	*2.5*	*45.6*	*1998*	*0.0*	*0.0*	*0.61*
Southern CIS	2.3	55.1	1995	14.6	4.1	0.58
Resource-poor	2.5	66.0	1995	21.9	5.1	0.53
Resource-rich	2.1	54.2	1996	7.4	3.3	0.55

Source: Fischer and Sahay (2000), 34.

Conclusions

Despite its size, level of development and degree of diversification, the Russian economy is a mineral economy and in line with the mineral-driven staple trap model, natural resource abundance did retard reform. The Russian government used the assumed surpluses of 'rich' energy companies to subsidise ailing industrial firms while borrowing future tax revenues from banks using state-owned energy enterprises as collateral. This brought temporary macro-economic stabilisation but institutionalised toleration of soft budget constraints and non-payment that then retarded reform. It also created a vicious circle of declining cash revenue and mounting debt that prevented long-term economic stabilisation and dictated continued high interest rates.

High interest rates rather than an inflow of foreign direct investment in mining drove the post-transition real appreciation of the rouble to double its equilibrium level, well above those of comparable transition economies other than mineral-rich Kazakstan. This depressed returns to domestic producers and reduced incentives for competitive restructuring. There is strong evidence of Dutch disease effects in Russia, as well as in oil-rich Azerbaijan and Kazakstan. The strong rebound of the real exchange rate contributed to a contraction in employment in the tradeables sectors to levels below those expected, despite the subsidies afforded to ailing SOEs by the soft budget constraints. The transfer of resources from potentially competitive enterprises to uncompetitive ones is consistent with the mineral-driven staple trap model in developing market economies. The resulting economic distortion is likely to lead to a growth collapse (Auty, 2001b).

Although rent-seeking was endemic by the time Russian reform commenced, and energy rents played a significant role, they were not the sole avenue. However, to the extent that energy subsidies were deployed in a way that achieved a stabilisation that proved temporary and meanwhile retarded reform and delayed the emergence of competitive markets, they fed the resurgence of rent-seeking behaviour by managers and officials in the mid-1990s. The rent-seeking occurred at the expense of tax revenues and social targeting so that income inequality in-

creased sharply and social capital was further eroded. Consequently, government efforts to cushion hardship by using energy rents to subsidise faltering SOEs and their employees failed. Any offsetting gains in terms of social welfare have been at best dubious over the medium-term. For example, Mosley and Kalyuzhnova (2000) estimate poverty expanded from 2 per cent to 50 per cent of the population during 1988–97. Moreover, the 'gains' are negative over the long term because the opportunity cost of the system lies in the sacrifice of an annual economic growth rate of 5–7 per cent that appears feasible if stabilisation is achieved and distortions to incentives are curbed.

This analysis suggests that policy should recognise more explicitly the consequences of Russia functioning as a mineral economy. Further external assistance should be made conditional upon a more transparent and rational use of its mineral rents. This requires not only economic policy reform but also the establishment of institutions to manage the rents. Such institutions include a windfall revenue tax to capture fluctuating rents and a capital development fund to sterilise them. A mineral revenue stabilisation fund is also required to smooth absorption of the rents into public expenditure, and a project evaluation unit to compare the returns to public investment. Finally, the model points to the paradoxical conclusion that both political and economic reform might be facilitated by a sustained decline in global energy prices.

Acknowledgement

Helpful comments from Mansoob Murshed are gratefully acknowledged. Any errors are, however, entirely those of the author.

Notes

1 However, although Johnson *et al*. (1997) agree that corruption is inversely related to social capital, they are reluctant to link corruption to natural resource abundance.
2 Interestingly, this implies that aid may prove counter-productive by providing the avenue for immediate personal enrichment that is normally associated with an abundance of natural resource rents.
3 Dutch disease effects are a loss of viability (often accompanied by a contraction in output) of the non-booming tradeable sectors that results from the boom-induced strengthening of the real exchange rate.
4 Although Markandya and Averchenkova (2001) estimate the subsidy at twice this figure and calculate the potential rents were twice as much again.
5 It seems likely that 'looting' in Russia involves foreign partnership in over-invoicing imports, money laundering, but also stock market scams. It also seems that in addition to 'white-collar' irregularities, Russia is at the leading edge of international crime involving arms sale to belligerents in African civil wars, smuggling people etc.

References

Åslund, A. (1999) 'Why has Russia's economic transformation been so arduous?', paper presented to the World Bank Annual Conference on Development Economics, Washington DC.

Auty, R.M. (ed.) (2001a) *Resource Abundance and Economic Development*, Oxford: Oxford University Press.

Auty, R.M. (2001b) 'Transition reform in the mineral-rich Caspian region countries', *Resources Policy* 27 (1): 25–32.

Baldwin, R.E. (1956) 'Patterns of development in newly settled regions', *Manchester School of Social and Economic Studies* 24: 161–79.

Bevan, D., P. Collier and J.W. Gunning (1987) 'Consequences of a commodity boom in a controlled economy: accumulation and redistribution in Kenya', *World Bank Economic Review* 1: 489–513.

Birdsall, N., T. Pinckney and R. Sabot (2001) 'Natural resources, human capital and growth', in Auty, R.M. (ed.), *Resource Abundance and Economic Development*, Oxford: Oxford University Press.

Bloom, D.E. and J.G. Williamson, (1998) 'Demographic transitions and economic miracles in emerging Asia', *The World Bank Economic Review* 12: 419–55.

Boone, P. and B. Fedorov (1997) 'The ups and downs of Russian economic reforms', in Woo, W.T., S. Parker and J.D. Sachs (eds.) *Economies in Transition: Comparing Asia and Europe*, Cambridge MA: MIT Press.

De Melo, M., C. Denizer, A. Gelb and S. Tenev (1997) 'Circumstance and choice: the role of initial conditions and policies in transition economies', mimeo, Washington DC: World Bank.

EBRD (1997) *Transition Report 1997: Enterprise Performance and Growth*, London: European Bank for Reconstruction and Development.

EBRD (1999) *Transition Report 1999: Ten Years of Transition*, London: European Bank for Reconstruction and Development.

EBRD (2000) *Employment Trends in Transition Economies 1989–98*, London: European Bank for Reconstruction and Development.

Economist (1999), 'Russian organised crime', *The Economist* August 29: 63–64.

Financial Times (1999a) 'Russia's oil groups dig deep to appease investors', *Financial Times* October 13.

Financial Times (1999b) 'Russia's oil groups find the rainbow', *Financial Times* June 2.

Fischer, S. and R. Sahay (1999) *The Transition Economies after Ten Years*, IMF Working Paper 00/30, Washington DC: IMF..

Gaddy, C.G. and B.W. Ickes (1998) 'Russia's virtual economy', *Foreign Affairs* 77 (5): 53–67.

Goldman, M. (1999) 'Russian energy: A blessing and a curse', *Journal of International Affairs* 53 (1): 73–84.

Gurgen, E., H. Snoek, J. Craig, J. McHigh, I. Izvorski and R. van Rooden (1999) *Economic Reforms in Kazakstan, Kyrgyz Republic, Tajikistan, Turkmenistan and Uzbekistan*, IMF Occasional Paper 183, Washington: IMF.

IMF (1999) 'Russian Federation: recent economic developments', *IMF Staff Country Report* 99/100, Washington DC: International Monetary Fund.

IMF (2000) *World Economic Outlook: October 2000*, Washington DC: International Monetary Fund.

Johnson, S., D. Kaufman and A. Shleifer (1997) 'The unofficial economy in transition', *Brookings Papers on Economic Activity* 2: 159–239.

Krueger, A.O. and B. Tuncer (1982) 'An empirical test of the infant industry argument', *American Economic Review* 72: 1142–52.

Leite, C. and J. Weidmann (1999) *Does Mother Nature Corrupt – Natural Resources, Corruption and Economic Growth*, IMF Working Paper 99/85, Washington DC. International Monetary Fund.

Londono, J.L. (1996) *Poverty, Inequality and Human Capital Development in Latin America, 1950–2025*, Washington DC: World Bank.

Markandya, A. and A. Averchenkova (2001) 'Reforming a large resource-abundant transition economy', in R.M. Auty, (ed.), *Resource Abundance and Economic Development*, Oxford: Oxford University Press.

Mauro, P. (1995) 'Corruption and growth', *Quarterly Journal of Economics* 90: 681–712.

Mosley, P. and V. Kalyuzhnova (2000) 'Are poverty and social goals attainable in the transition region?' *Development Policy Review* 18: 107–20.

Murphy, K., A. Shleifer and R. Vishny (1991) 'The allocation of talent: implications for growth', *Quarterly Journal of Economics*, 106: 503–30.

Murphy, K., A. Shleifer and R. Vishny (1993) 'Why is rent seeking so costly to growth', *American Economic Review*, 83: 409–14.

Murshed, S.M. (2001) 'Short-run models of contrasting natural resource endowments', in R.M. Auty (ed.) *Resource Abundance and Economic Development*, Oxford: Oxford University Press.

Nankani, G. (1979) *Development Problems of Mineral Exporting Countries*, World Bank Staff Working Paper 354, Washington DC: World Bank.

Pinto, B., V. Drebentsov and A. Morozov (1999) *Dismantling Russia's Non-payment System: Creating Conditions for Growth*, Moscow: World Bank.

Rosenberg, C.B. and T.O. Saavalainen (1998) *How to Deal with Azerbaijan's Oil Boom? Policy Strategies in a Resource-rich Transition Economy*, IMF Working Paper 98/6, Washington DC: IMF.

Sachs, J.D. (1999) 'Resource endowments and the real exchange rate: a comparison of Latin America and East Asia', in Ito, T. and A.O. Krueger (eds.) *Changes in Exchange Rates in rapidly Developing Countries*, Chicago: University of Chicago Press.

Sachs, J.D. and A.M. Warner (1995) 'Economic reform and the process of global integration', *Brookings Papers on Economic Activity* 1: 1–118.

Shleifer, A. and D. Treisman (2000) *Without a Map: Political Tactics and Economic Reform in Russia*, Cambridge MA: MIT.

Wood, A. and K. Berge (1997) 'Exporting manufactures: human resources, natural resources, and trade policy', *Journal of Development Studies* 34: 35–59.

Woolcock, M., J. Isham and L. Pritchett (2001) 'The social foundations of poor economic growth in resource-rich countries, in R.M. Auty (ed.) *Resource Abundance and Economic Development*, Oxford: Oxford University Press.

World Bank (1999) *World Development Indicators*, Washington DC: World Bank.

World Bank (2000) *World Development Indicators*, Washington DC: World Bank.

WRI (1998) *World Resources 1998–99: Environmental Change and Human Health*, New York: Oxford University Press. Chapter 4 of Issues in Positive Political Economy, MURSHED (ed), by Auty 11
4-12

5 Electoral uncertainty, economic policy and growth

George Economides, Natasha Miaouli and Apostolis Philippopoulos

It is widely believed that political factors (elections, partisan motives, and bureaucracy) are crucial in determining economic policies and in turn economic outcomes.[1] The early literature based on rational expectations focused on the link between elections and fiscal policy. For instance, Rogoff and Sibert (1988) showed how the incumbent political party manipulates policy instruments in an attempt to increase its re-election probability. Persson and Svensson (1989) and Alesina and Tabellini (1990) developed two-party models to show how fiscal and public debt policy can be used strategically by the incumbent party to influence the choices of its successor. In Lockwood *et al*. (1996), electoral uncertainty reduces the marginal cost of public debt and this leads to relatively loose fiscal policy before elections, in the form of over-spending, under-taxing and over-borrowing.

More recently, the emphasis has been on economic growth. There is robust empirical evidence that socio-political instability affects economic growth. This is a rich literature that includes Barro (1991), Easterly and Rebelo (1993), Barro (1996), Levine and Zervos (1996), Alesina and Perotti (1996), Alesina *et al*. (1997), Rodrik (1997), Devereux and Wen (1998), Darby *et al*. (1998) and many others (for a recent survey, see Drazen (2000), chapter 11). Most of these papers use various sociopolitical indices (measures of democracy, political violence, government duration, and income inequality) in *ad hoc* growth regressions to see how socio-political factors affect economic growth.

The present chapter formalizes the link among electoral uncertainty (in the form of re-election probabilities), fiscal policy and economic growth. To do so, it builds upon previous work by Alesina and Tabellini (1990), Lockwood *et al*. (1996), Devereux and Wen (1998), Persson and Tabellini (1999a) and Asteriou *et al*. (2000).[2] The setup is a two-party, general equilibrium model of optimal growth and fiscal policy, in which the elected party chooses distorting taxes to finance government consumption expenditures. We use a fully dynamic infinite-time horizon model, so that we do not ignore any important dynamic implications. Within this politico-economic setup, we study how re-election probabilities (interpreted as a measure of political uncertainty) affect the conduct of fiscal policy, and in turn how fiscal policy affects private investment and economic growth.

The model is as follows. We consider a closed economy with a private sector (households and firms) and two political parties that can alternate in power because of elections. Households consume, work and save in the form of capital. Firms use capital and labour to produce a single good. The elected political party forms a government that finances its public consumption services by taxing households' income. The dynamic way we model the electoral system is similar to that in Lockwood *et al*. (1996). That is, the elected party chooses economic policy while in office by playing Stackelberg *vis-à-vis* private agents and Nash *vis-à-vis* the other party, which may regain power in the next election with a non-zero probability. Since optimal tax policies are inherently time-inconsistent, we solve for Markov strategies, and hence a Markov-perfect general equilibrium in which optimal policies are time-consistent.[3] We work as follows: we first solve for a competitive equilibrium, given any (Markov) fiscal policy; then, we endogenize fiscal policy by solving for Markov strategies.

There are two main results for economic policy. First, the optimal income tax rate is flat over time. This is a version of Barro's (1979) tax smoothing result: it is optimal to smooth out tax distortions over time. In other words, it is optimal for policymakers to keep the tax rate (and the associated government expenditures-to-output ratio) constant over time, even if the underlying general equilibrium model is non-linear and economic policy is the outcome of a game between two political parties that alternate in power, rather than the choice of a single benevolent government. Second, the optimal income tax rate (and the associated government expenditures-to-output ratio) decreases with the probability of getting re-elected. In other words, when re-election becomes less certain, policymakers find it optimal to go for a larger public sector. In turn, the higher tax rate – required to finance the larger public sector – reduces private capital accumulation and economic growth. In other words, electoral uncertainty can distort the size and role of government. This is consistent with empirical evidence of a negative effect of 'too large' government sizes upon economic growth in OECD economies (see Tanzi and Schuknecht, 1997, 2000).

These theoretical results are similar to those in Alesina and Tabellini (1990) Lockwood *et al.* (1996), Devereux and Wen (1998), Persson and Tebellini (1999a) and Asteriou *et al.* (2000). Asteriou *et al*. (2000) also provide direct empirical support for our results. Specifically, by using UK data, they confirm that lower *ex ante* re-election probabilities (calculated by using opinion polls) lead to lower economic growth. Note that since they use the incumbent's popularity as a measure of *ex ante* re-election probabilities, they provide evidence different from that of the literature that has mostly used ad hoc indices of sociopolitical instability.

The mechanism that drives our results is as follows. When there is electoral uncertainty (in the sense that there is a non-zero probability of being out of power in the next election) and the political parties do not care enough about economic outcomes when out of power (specifically, we need to assume that the parties care about economic outcomes less when out of power),[4] they effectively face a *quasi finite time-horizon*. As a result, the lower the probability of getting re-elected, the smaller the effective discount rate, the less the incumbent party values capital accumulation, and the more it spends on current unproductive activities. In other words,

electoral uncertainty induces policymakers to follow shortsighted, inefficient policies; here, the inefficiency takes the form of a relatively large government sector with short-term benefits, high tax burden and eventually low economic growth. Note that, in general, one needs more than one imperfection to generate such inefficiency. Here, it is electoral uncertainty in combination with the assumption that political parties care relatively little about economic outcomes when out of power (see also Economides *et al*. (2001)).

The rest of the chapter is as follows. The first section presents the theory. The second section reviews the empirical literature. The third section concludes with a general discussion.

A politico-economic model of optimal growth

Consider a closed economy with a private sector and two political parties. The private sector consists of a representative household and a representative firm. The household consumes, works and saves in the form of capital. The firm produces a single good by using labour and capital. The political party in power forms a government that finances public consumption services by taxing the household's income. All economic policy instruments are endogenous, i.e. optimally chosen. We assume discrete time and infinite time-horizons. We will first solve the private agents' optimization problems, for any given economic policy. Then, economic policy will be determined by a Nash game between two alternating political parties.

Households

The representative household maximizes intertemporal utility:

$$\sum_{t=0}^{\infty} \beta^t u(c_t, g_t) \tag{5.1a}$$

Where c_t and g_t are respectively private consumption and government consumption at time t, and the parameter $0 < \beta < 1$ is the discount rate. The instantaneous utility function is increasing and concave, and satisfies the Inada conditions. For algebraic simplicity, we assume that $u(\cdot)$ is additively separable and logarithmic. Thus,

$$u(c_t, g_t) = \log c_t + \delta \log g_t \tag{5.1b}$$

Where the parameter $\delta \geq 0$ is the weight given to government consumption relative to private consumption.

The household rents its beginning-of period capital, k_t, to the firm and receives $r_t k_t$, where r_t denotes the return to capital at t. It also supplies inelastically one unit of labour services per unit of time and receives labour income, w_t.[5] Further, it receives profits, π_t. Thus, the household's flow budget constraint is:

$$c_t + k_{t+1} = (1 - \theta_t)(r_t k_t + w_t + \pi_t) \tag{5.2}$$

Where $0 < \theta_t < 1$ is the income tax rate. The initial capital stock, k_0, is given. For algebraic simplicity, we assume full capital depreciation.

The household acts competitively by taking prices, tax policy and public services as given. From the household's viewpoint, the state at any time t can be summarized by the predetermined capital stock, k_t, and the current tax rate, θ_t. We will solve this optimization problem by using the technique of dynamic programming. Then, let $V(k_t;\theta_t)$ denote the value function of the household at time t . This value function must satisfy the Bellman equation:

$$V(k_t;\theta_t) \equiv \max_{c_t,k_{t+1}} [\log c_t + \delta \log g_t + \beta V(k_{t+1};\theta_{t+1})] \tag{5.3}$$

Using (5.2) into (5.3) for c_t, the optimality condition for k_{t+1} is:

$$\frac{1}{c_t} = \beta V_k(k_{t+1};\theta_{t+1}) \tag{5.4a}$$

while the envelope condition for k_t is:[6]

$$V_k(k_t;\theta_t) = \frac{(1-\theta_t)r_t}{c_t} \tag{5.4b}$$

Firms

Technology takes a Cobb–Douglas form. The production function is:[7]

$$y_t = Ak_t^a \tag{5.5}$$

where $A > 0$ and $0 < \alpha < 1$ are parameters.

At any time t, the representative firm maximizes profits, π_t:

$$\pi_t \equiv y_t - r_t k_t - w_t \tag{5.6}$$

The firm acts competitively by taking prices as given. This is a static problem. The standard first-order conditions, that also imply zero profits, are:

$$r_t = aAk_t^{a-1} \tag{5.7a}$$

$$w_t = (1-a)Ak_t^a \tag{5.7b}$$

Government budget constraint

We assume that at each time t, the government runs a balanced budget, and for simplicity there is no public debt. That is, the government finances public consumption services, g_t, by taxing the household's income at a rate θ_t. The government's flow budget constraint is:

$$g_t = \theta_t(r_t k_t + w_t + \pi_t) \tag{5.8}$$

Competitive equilibrium (given economic policy)

Given tax policy, $\{\theta_t\}_{t=0}^{\infty}$, a competitive equilibrium (CE) is defined to be a sequence of allocations $\{c_t, k_{t+1}, g_t\}_{t=0}^{\infty}$, and prices $\{r_t, w_t\}_{t=0}^{\infty}$, such that:[8] (i) households maximize utility and firms maximize profits, given prices and economic policy; (ii) all budget constraints are satisfied; (iii) all markets clear via price flexibility. This CE is characterized by equations (5.1)–(5.8) above. In this subsection, we will take advantage of the specific functional forms used to get a closed-form analytical solution for this CE.

We start with economy-wide output. Equations (5.5), (5.6) and (5.7a)–(5.7b) imply:

$$r_t k_t + w_t + \pi_t = y_t = A k_t^a \tag{5.9}$$

By making use of the Cobb–Douglas constraint in (5.9) and the log-linear utility function in (5.1a)–(5.1b), equations (5.1)–(58) imply:[9]

Result 5.1 *In a competitive equilibrium (given any Markov economic policy), optimal private consumption and end-of-period capital follow:*

$$c_t = A(1 - a\beta)(1 - \theta_t)k_t^a \tag{5.10a}$$

$$k_{t+1} = Aa\beta(1 - \theta_t)k_t^a \tag{5.10b}$$

Note two things in (5.10a)–(5.10b). First, these are closed-form analytic solutions. This is thanks to the special structure of the model: log-linear utility functions, Cobb–Douglas production functions and full capital depreciation.[10] Second, the sign of $\frac{\partial c_t}{\partial \theta_t}$ and $\frac{\partial k_{t+1}}{\partial \theta_t}$ is always negative. Thus, private consumption and capital accumulation decrease with the current tax rate, θ_t (see Asteriou *et al.*, 2000, for a richer model in which the sign of $\frac{\partial k_{t+1}}{\partial \theta_t}$ is non-monotonic).

It will be useful for what follows, to present also the government's budget constraint in a CE. Using (5.9) into (5.8), government expenditures are simply:

$$g_t = \theta_t A k_t^a \tag{5.10c}$$

so that the expenditures-to-output ratio will be constant, if the tax rate is constant (see below).

We summarize the findings so far. Equations (5.10a), (5.10b) and (5.10c) give c_t, k_{t+1} and g_t respectively in a competitive equilibrium (CE). This is for any Markov fiscal policy, where the latter is summarized by the current tax rate, θ_t. Note that c_t, k_{t+1} and g_t are functions of the predetermined capital stock, k_t, and the current tax rate, θ_t, only. This will make the political parties' optimization problem recursive and hence optimal policies will be time consistent.[11] The next subsection will endogenize the choice of θ_t.

Definition of political general equilibrium

To endogenize economic policy, we form a Nash game between two political parties, denoted by *i* and *j*, which can alternate in power according to an exogenous

re-election probability,$0 \leq q \leq 1$. For simplicity, elections take place in each time period.[12] Thus, the party in power at time t has a probability q of winning the next election and remaining in power in the next time period $t+1$, and a probability $1-q$ of losing the election and remaining out of power at $t+1$. The elected party chooses the current tax rate, θ_t, to maximize the utility of the representative household. In doing so, it plays Stackelberg *vis-à-vis* private agents (households and firms). It also plays Nash *vis-à-vis* the other political party, which may be in power in the next time period with a non-zero probability. Specifically, the political general equilibrium (PGE) is defined as follows:

1. In each time-period t, the elected party i chooses its tax policy to maximize (5.1a)–(5.1b) subject to the competitive equilibrium (summarized by (5.10a)–(5.10c) above) and by taking as given the policy of the other party, $j \neq i$, which may be in power at $t+1$.
2. We solve for symmetric Nash strategies. That is, since the two parties are assumed to be alike, their (Nash) strategies are symmetric *ex post*.[13]
3. The parties do not care about economic outcomes when out of power. This assumption is made for simplicity. All we need to assume is that the political parties care about economic outcomes less when out of power than when in power (see below for details).
4. We focus on Markov policy strategies. That is, θ_t will be a function of the current value of the economy-wide state variables only. Note that this also confirms the solution to the private agents' optimization problem in the previous subsection (see Result 5.1 above).
5. The solution for θ_t, in combination with the CE above, will give a Markov-perfect general equilibrium, which we call political general equilibrium (PGE). This equilibrium is similar to that in Asteriou *et al*. (2000) and Economides *et al*. (2001).

Problem formulation

We solve the above problem by using the technique of dynamic programming. From the political parties' viewpoint, the state variable at any time t is the economy's inherited stock of capital, k_t. Let $V^{P_i}(k_t)$ and $V^{N_i}(k_t)$ be respectively the value functions of party i when in power, and when out of power, at time t. Then, $V^{P_i}(k_t)$ and $V^{N_i}(k_t)$ must satisfy the following pair of Bellman equations:[14]

$$V^{P_i}(k_t) = \max_{\theta_t} \left[\log c_t + \delta \log g_t + \beta q V^{P_i}(k_{t+1}) + \beta(1-q)V^{N_i}(k_{t+1})\right] \tag{5.11a}$$

$$V^{N_i}(k_t) = 0 + \beta(1-q)V^{P_i}(k_{t+1}) + \beta q V^{N_i}(k_{t+1}) \tag{5.11b}$$

where c_t, k_{t+1} and g_t follow (5.10a), (5.10b) and (5.10c) respectively. Notice that in (5.11a), the incumbent party has a probability q of remaining in power and a

probability $1-q$ of losing the coming election. In (5.11b), when the party is out of power, it knows that there is a probability q of continuing to be out of power and a probability $1-q$ of coming back to power in the next election. Also notice that when out of power, the parties do not care about policy outcomes; hence the zero in (5.11b).

Optimal policy and political general equilibrium

Solution of the policy problem in (5.11a)–(5.11b) implies:[15]

Result 5.2 *In a Political General Equilibrium, the optimal income tax rate, θ_t, is constant over time and equal to:*

$$0 < \theta = \frac{\delta}{\delta + \Omega} < 1 \tag{5.12}$$

where,

$$\Omega = 1 + \beta\left(qu_1^P + (1-q)u_1^N\right) = \frac{(1-a\beta)\left(1+a\beta(1-2q)\right) + a\beta(1+\delta)\left(q+a\beta(1-2q)\right)}{(1-a\beta)\left(1+a\beta(1-2q)\right)}$$

Notice two things in (5.12). Firstly, it is optimal to keep the policy instrument flat over time. This is a tax smoothing result. Obviously, this type of policy introduces fewer intertemporal distortions. Second, the 'effective' discount rate, Ω, increases with the probability of getting re-elected, q, i.e. $\frac{\partial \Omega}{\partial q} > 0$. That is, as the probability of getting re-elected increases, policymakers care effectively more about the future. In turn, (5.12) implies $\frac{\partial \theta}{\partial q} < 0$. That is, as the probability of getting re-elected increases, the tax rate – and the associated government expenditures-to-output ratio – decreases. Since c_t and k_{t+1} decrease with the tax rate θ_t (see equations (5.10a) and (5.10b) above), it follows that as the probability of remaining in power increases, both consumption and growth increase.

Therefore, when (a) there is electoral uncertainty and (b) the parties care relatively little about economic outcomes when out of power, they effectively face a quasi-finite time-horizon. In this case, as the re-election probability gets smaller (i.e. q falls), the party in power cares less about the future and spends more now. Since higher spending requires higher tax revenues, the tax rate is higher than without electoral uncertainty.[16] This is bad for growth.

These results are summarized in the following proposition:

Proposition 5.1 *There is a unique Markov-perfect general equilibrium in Nash strategies among political parties. In this political equilibrium, when the probability of getting re-elected decreases, it is optimal for policymakers to follow short-sighted fiscal policies (in the form of a higher total expenditure-to-output ratio) and this is bad for private capital accumulation and economic growth.*

Note that the above result can also hold in richer setups with more economic policy instruments. In particular, when Asteriou *et al*. (2000) include government

production services that provide Barro-type production externalities to private firms, they show that the share of tax revenues allocated to government production services *vis-à-vis* government consumption services increases with the re-election probability. In other words, when political uncertainty increases (i.e. the re-election probability falls), not only the tax rate increases as in the present model, but also the allocation of total tax revenues turns in favour of non-productive activities. Both economic policy instruments work in the same direction, so that electoral uncertainty again reduces economic growth.

It is important to emphasize that the main result of the literature (i.e. the lower the re-election probability, the stronger the incentive for policymakers to choose short-sighted distorted fiscal policies, and hence the lower is economic growth and intertemporal welfare) holds under the crucial assumption that political parties care about economic outcomes more when in than out of power (see the Bellman equations (5.11a)–(5.11b) above). This point has been made clear by Lockwood *et al*. (1996). Using a more general setup, Economides *et al*. (2001) revisit the underlying mechanism and confirm that to get this realistic result, re-election uncertainty must indeed be combined with the hypothesis that political parties care about economic outcomes more when in than out of power.[17] More importantly, they show that this preference over being in power is *ad hoc* (exogenously set). In other words, if the incumbent party can also choose how much to care about economic outcomes when in and out of power, it is optimal to care the same so that short-sighted distorted policies, like the one derived above, do not arise.

Empirical evidence: a review of the literature

This section will briefly review the relevant empirical literature and summarize its methodology, main findings and problems. Following the theoretical analysis above, we will focus on the link among political uncertainty, fiscal policy and economic growth.

The connection between fiscal policy and economic growth has long been a central area of empirical research. Recently, thanks mainly to the theory of endogenous growth, interest has focused on the implications of the different tax structures and the different types of government expenditures for economic growth. For instance, Mendoza *et al*. (1997) provide evidence that distortionary taxation harms investment and economic growth, while non-distortionary taxation does not. Kneller *et al*. (1998, 1999) provide similar evidence for taxes and also find that productive government expenditures enhance growth, while non-productive expenditures harm growth (see Asteriou *et al*. (2000) for a recent survey).

Also, since the early 1960s in most OECD countries, there is evidence that: (i) the size of the public sector (measured by total government expenditures as a share of GDP) has increased substantially; and (ii) government consumption (for example, transfers and government wages) as a share of GDP shows a sharply upward movement relative to government investment (see Alesina, 1999, and Tanzi and Schuknecht, 2000). Thus, there seems to be a negative relationship between large-sized governments and economic and social performance (see Tanzi and

Schuknecht, 1997, 2000). These trends in the data can be explained by theories of political distortions. For instance, as we showed above, electoral uncertainty can induce policymakers to go for short-sighted policies, in the form of too large public sectors and biases in favour of unproductive activities, and this is bad for economic growth.

One is then led to investigate the empirical link among political distortions, economic policy instruments and economic growth. To do so, the literature has used several indices of political distortions. For instance, inequality in the distribution of income; socio-political instability in the form of frequent government turnovers, political violence and unrest; political corruption, bureaucracy and lack of property rights; pre-electoral bribes and fiscal euphoria. Although there are differences across different studies, two results seem to be rather robust: (a) there is a correlation between political distortions and the conduct/manipulation of economic policy; (b) there is a negative effect from political distortions to economic growth, see Alesina *et al*. (1997), Alesina (1999) and Drazen (2000). Note that the literature has emphasized not only how policies affect the macro-economy but also how variability (i.e. lack of persistence) in policies affects the macro-economy.

However, several empirical issues are still open. As Drazen (2000, chapter 11) points out and our theoretical model above makes clear, one should be able to divide the reduced-form effect from political distortions to economic outcomes into two sub-effects: from distortions to economic policy instruments, and in turn from economic policy instruments to economic outcomes. Unfortunately, it is rather rare to find significant support for both sub-effects simultaneously, especially if one tests formally the cross-equation restrictions implied by the (general equilibrium) theory. This is why most empirical studies have relied exclusively on unrestricted reduced-form regressions. Asteriou *et al*. (2000) is a recent example of this methodology in a theoretical model similar to the one presented above. When they use government's popularity as a measure of *ex ante* re-election probabilities, they find a strong negative effect from low popularity to economic growth in the UK. However, the data cannot distinguish the two sub-effects (i.e. from popularity to fiscal policy instruments, and from fiscal policy instruments to economic growth). We believe this is a promising area for further empirical research.

Conclusions and related politico-economy issues

In this chapter, we have used a general equilibrium politico-economy model to formalize the link among electoral uncertainty, fiscal policy and economic growth. We showed that lower re-election probabilities can create pressure for short-sighted fiscal policies, and this is bad for economic growth. This result follows from a combination of electoral uncertainty and the assumption that the parties care relatively little about economic outcomes when out of power. These predictions are consistent with empirical evidence.

So far we have been rather formal. We will end the chapter with a discussion of some closely related politico-economy issues.

Market failures versus policy failures

Although electoral uncertainty can induce short-sighted fiscal policies, this is obviously not a message against elections or multi-party political systems. It is well known that elections and political parties have a multiple role to play within a society. For instance, as Persson and Tabellini (1994) point out, elections make officeholders accountable to the electorate, in the sense that voters have the chance to select either the most competent policymaker or the policymaker whose ideology is closer to the majority of the voters. Elections can also control the moral hazard of the policymakers, who have to limit their opportunism to reduce the threat of replacement.

Instead, the way to interpret our results is as another application of the tradeoff between 'market failures' and 'policy failures'. The general idea is that when there are market failures (public goods, externalities, monopolistic situations, influential lobbies and so on), policy intervention is needed to correct these failures, and fiscal policy is an obvious candidate for this role. It is then the tradeoff between market and policy failures that determines the optimal policy. What we have shown in this chapter (as well as in our companion papers in Asteriou *et al*., 2000 and Economides *et al*., 2001) is that electoral uncertainty pushes policymakers (who care relatively little about economic outcomes when out of power) to follow short-sighted inefficient policies.[18] Therefore, while economic policy is needed to correct the existing market failures, it also generates its own distortions.[19]

Need for enforcement mechanisms

Given the above, a natural question to ask is the following: can we design mechanisms that reduce political distortions? In the context of our multi-party democratic model, the task is to design mechanisms that give policymakers the incentive to care about economic outcomes all the time, and not just when they happen to be in power. In Rogoff (1990) type models, the task is to design mechanisms that discourage policymakers from manipulating pre-election probabilities. In Alesina and Drazen (1991) type models, the task is to design mechanisms that force policymakers to stop the war of attrition and break the vicious cycle of the status quo.

Therefore, in most politico-economy models, the main task is to design mechanisms that can extend the effective time horizon of policymakers, without harming the function of multi-party democracies. Note that this is consistent with the general game theoretic result that many properties, concerning the comparison between non-cooperative and cooperative outcomes, may change once the model allows for dynamics (recall that the one-shot equilibrium of the prisoner's dilemma game is not necessarily the equilibrium of the same game repeated a sufficiently large number of times). It is also consistent with the literature on public goods. For instance, Glomm and Lagunoff (1999) show that, concerning the provision of public goods, whether voluntary (i.e. decentralized) or coercive (i.e. centralized) mechanisms prevail, depends crucially on whether the game is static or dynamic. This is because, in a dynamic set-up, the accumulation process mitigates the prob-

lem of conflicting interests occurring in coercive communities and hence such communities become more attractive.

How can we force policymakers to care about the future? The ideal scenario would be a kind of endogenous enforcement. However, this is too good to be true in a world that typically exhibits political distortions, and where elections have usually the features of a one-shot noncooperative game. In this case, the task is to establish a constitutional and political system (through monitoring, provision of information to the public, fines/subsidies, credible punishment of political corruption, etc) that increases political competition between selfish politicians and gives them incentives for more far-sighted policies. There is a rapidly developing literature on these issues that are beyond the scope of this discussion (see Laffont (1999), Myerson (1999) and Persson and Tabellini (1999b)).

All this means that we do not only need game theory but also implementation theory. As Moore (1992: 188–9) has pointed out, the former is concerned with how a given game is played, while the latter is concerned with the design of the game. The choice of mechanism will be driven by the choice of the equilibrium concept. Our preference should be for a cooperative equilibrium concept in general, or (in the context of our model) a mechanism in which the political parties care about economic outcomes the entire time even when they are out of power. In other words, a complete theory on the link between politics and economics requires an analysis of both the 'electoral' and the 'governmental' level (see Myerson (1999: 672–3)). As Myerson says, the electoral level defines the procedures, by which candidates are elected, while the governmental level defines the channels through which the constitutional structure of the political system affects the way economic policy is formed. In the present chapter, as in most of the relevant literature, we focused only on one of these levels; on the governmental level, taking as given the electoral system. To sum up, the two-way link between politics and economics is expected to remain one of the main determinants of the way economic policy is conducted. The need for setting up the right incentives for policymakers is a big challenge in democratic societies.

Acknowledgements

We thank Tryphon Kollintzas, Sajal Lahiri, Ben Lockwood, Jim Malley, Thomas Moutos, S. Mansoob Murshed, Simon Price, Euclid Tsakalotos, Panos Tsakloglou and Vangelis Vassilatos for comments and discussions. We also thank seminar participants at the IESG Easter mini-conference, University of Warwick, April 1999. All errors are ours.

Notes

1 For a recent survey of the literature on politics and macroeconomics, see Drazen (2000).
2 See our companion papers Asteriou *et al*. (2000) and Economides *et al*. (2001) for a comparison of these models. See also Persson and Tabellini (1999a) for a survey of the literature on political uncertainty, economic policy and economic growth.

3 Thus, optimal policies depend on the current value of the relevant state variables. For the properties of Markov strategies and Markov-perfect equilibria in macroeconomic setups see Obstfeld (1991).
4 Laver and Hunt (1992) provide evidence from the political science literature that this is indeed the case in most democratic countries. See Lockwood *et al*. (1996) for details.
5 Assuming inelastic labour supply does not affect our main results.
6 See also Sargent (1987, chapter 1). Equations (5.4a)–(5.4b) give the standard Euler equation, i.e.

$$\frac{c_{t+1}}{c_t} = \beta(1 - \theta_{t+1} r_{t+1})$$

.
7 The firm's problem is written in labour intensive form. Then, in equilibrium, there is one unit of labour services. For details, see Barro and Sala-i-Martin (1995, chapter 2).
8 Thus, we choose to solve for a CE in terms of tax rates $\{\theta_t\}_{t=0}^{\infty}$. Alternatively, we could solve in terms of spending $\{g_t\}_{t=0}^{\infty}$. Below, we endogenize the choice of $\{\theta_t\}_{t=0}^{\infty}$.
9 We work as follows: In a CE, the structure of the problem implies a conjecture of the value function in equation (5.3) of the form $V(k_t;\theta_t) = u_0 + u_1 \log k_t + u_1 \theta_t + u_3 \log \theta_t$, where u_0, u_1, u_2, u_3 are undetermined coefficients. Then, the optimality conditions (5.4a)–(5.4b) together with equations (5.5), (5.6) and (5.7a)–(5.7b) give (5.10b). In turn, (5.10a) follows from (5.10b) and the budget constraint (5.2). Then, plugging (5.10a)–(5.10b) back into (5.3) and equating coefficients on both sides of the Bellman, we can solve for u_0, u_1, u_2, u_3. Note that while we can solve for u_1 at this stage, we cannot solve for u_0, u_2, u_3 before we also solve for optimal policy in the next section. This is how it should be in a general equilibrium model where policy is endogenously chosen. For details, see Asteriou *et al*. (2000) and Malley *et al*. (2001).
10 See also Sargent (1987), Stokey and Lucas (1989) and Obstfeld and Rogoff (1996). For a similar model with a single government, see Malley *et al*. (2001).
11 See Obstfeld (1991). For similar applications and details, see Kollintzas *et al*. (2000) and Malley *et al*. (2001).
12 See Lockwood *et al*. (1996) for a richer model in which the electoral cycle lasts two time periods so that the elected party can remain in power for two periods. Our main results do not depend on this.
13 See Lockwood *et al*. (1996) for partisan differences in a public finance model.
14 This modelling is borrowed from Alesina and Tabellini (1990) and Lockwood *et al*. (1996).
15 We work as follows: inspection reveals that we have to solve a dynamic programming problem with a log-linear payoff function and Cobb–Douglas constraints. Thus, the functional formulation of the policymakers' problem is similar to that of the private agents' above. This means that the value functions in (5.11a)–(5.11b) are expected to be of the log-linear form $V^P(k_t) = u_0^P + u_1^P \log k_t$ and $V^N(k_t) = u_0^N + u_1^N \log k_t$, where $u_0^P, u_1^P, u_0^N, u_1^N$ are undetermined coefficients. Using this conjecture for the value functions into (5.11a)–(5.11b), differentiating the right-hand side of (5.11a) with respect to θ_t, imposing the symmetricity conditions $\theta_t^i = \theta_t^j \equiv \theta_t$, $u^{Pi} = u^{Pj} \equiv u^P$ and $u^{Ni} = u^{Nj} \equiv u^N$, plugging the optimality conditions back into (5.11a)–(5.11b) and equating coefficients on both sides of the Bellman equations, we solve for $u_0^P, u_1^P, u_0^N, u_1^N$. Note that this also completes the solution for the CE above. For details, see Asteriou *et al*. (2000) and Malley *et al*. (2001).
16 When $q = 1$, we get the benevolent government case (second best).
17 That is, Economides *et al*. (2001) show that if there is electoral uncertainty (i.e. $q < 1$) but the parties care the same about economic outcomes all the time irrespectively of whether they are in power or not, then the solution is equivalent to that of a benevolent (one party) government.
18 There is a view that problems like these will possibly disappear in the near future. As Alesina *et al*. (1997) point out, this is based on two arguments: (a) the EU integration

process will make national policies much less independent, and (b) many countries will soon face problems of oversized public sectors. Although both (a) and (b) are true, they are not enough to eliminate political business cycles. Even within the EU, member-countries will not completely lose their fiscal autonomy. Also, globalization may worsen some of the existing political distortions. For instance, within the new integrated environment, national governments have a stronger incentive to free ride on other economies (an example is EU redistributive transfers). Also, when unpopular measures of fiscal stabilization are taken, this can increase political distortions in the short run (rent seeking activities). See Alesina *et al*. (1997) for these issues

19 In addition to shortsighted policies due to electoral uncertainty, other popular policy distortions are manipulation of asymmetric information, transfers to powerful groups of voters, rent-seeking bureaucracies, political corruption, etc.

References

Alesina, A. (1999) 'Too large and too small governments', in V. Tanzi, K. Chu and S. Gupta (eds) *Economic Policy and Equity*, Washington DC: International Monetary Fund.

Alesina, A. and A. Drazen (1991) 'Why are stabilizations delayed?', *American Economic Review*, 81: 1170–88.

Alesina, A. and R. Perotti (1996) 'Income distribution, political instability and investment', *European Economic Review*, 40: 1203–28.

Alesina, A. and G. Tabellini (1990) 'A positive theory of fiscal deficits and government debt', *Review of Economic Studies*, 57: 403–14.

Alesina, A., N. Roubini and G. Cohen (1997) *Political Cycles and the Macroeconomy*. The MIT Press.

Asteriou, D., G. Economides, A. Philippopoulos and S. Price (2000) *Electoral Uncertainty, Fiscal Policy and Economic Growth: Theory and Evidence From the UK and a Panel of Parliamentary Democracies*, Discussion Paper, no. 00–11, Athens: Athens University of Economics and Business.

Barro, R. (1979) 'On the determination of public debt', *Journal of Political Economy*, 87: 940–71.

Barro, R. (1991) 'Economic growth in a cross-section of countries', *Quarterly Journal of Economics*, 106: 407–43.

Barro, R. (1996) 'Democracy and growth', *Journal of Economic Growth*, 1: 1–27.

Barro, R. and X. Sala-i-Martin (1995) *Economic Growth*. New York: McGraw Hill.

Darby, J., C. Li and A. Muscatelli (1998) 'Political uncertainty, public expenditure and growth', Discussion Paper, no. 9822, Glasgow: University of Glasgow.

Devereux, M. and J.-F. Wen (1998) 'Political instability, capital taxation and growth', *European Economic Review*, 42: 1635–51.

Drazen, A. (2000): *Political Economy in Macroeconomics*. Princeton: Princeton University Press.

Easterly, W. and S. Rebelo (1993) 'Fiscal policy and economic growth: an empirical investigation', *Journal of Monetary Economics*, 32: 417–58.

Economides, G., A. Philippopoulos and S. Price (2001) *Elections, Fiscal Policy and Growth: revisiting the Mechanism*, Discussion Paper no. 5, Athens: Athens University of Economics and Business.

Glomm, G. and R. Lagunoff (1999) 'A dynamic Tiebout theory of voluntary vs. involuntary provision of public goods', *Review of Economic Studies*, 66: 659–77.

Kneller, R., M.F. Bleaney and N. Gemmel (1998) *Growth, Public Policy and the Government Budget Constraint: Evidence from OECD Countries*, Discussion Paper, no. 98/14, Nottingham: University of Nottingham.

Kneller, R., M. Bleaney and N. Gemmell (1999) 'Fiscal policy and growth: evidence from OECD countries', *Journal of Public Economics*, 74: 171–90.

Kollintzas, T., A. Philippopoulos and V. Vassilatos (2000) 'Is tax policy coordination necessary?', Discussion Paper, no. 2348, Center for Economic Policy Research, London: CEPR.

Laffont, J.-J. (1999) 'Political economy, information and incentives', *European Economic Review*, 43: 649–69.

Laver, M. and B. Hunt (1992) *Policy and Party Competition*. London: Routledge.

Levine, R. and S. Zervos (1996) 'Stock market development and long-run growth', *The World Bank Economic Review*, 10: 323–39.

Lockwood, B., A. Philippopoulos and A. Snell (1996) 'Fiscal policy, public debt stabilization and politics: theory and UK evidence', *Economic Journal*, 106: 894–911.

Malley, J., A. Philippopoulos and G. Economides (2001) 'Testing for tax smoothing in a general equilibrium model of endogenous growth', working paper, forthcoming in *European Journal of Political Economy*.

Mendoza, E., G. Milesi-Ferretti and P. Asea (1997) 'On the effectiveness of tax policy in altering long-run growth: Harberger's superneutrality conjecture', *Journal of Public Economics*, 66: 99–126.

Moore, J. (1992) 'Implementation, contracts and renegotiation in environments with complete information', in J.-J. Laffont (ed.) *Advances in Economic Theory*, Cambridge: Cambridge University Press.

Myerson, R. (1999) 'Theoretical comparisons of electoral systems', *European Economic Review*, 43: 671–97.

Obstfeld, M. (1991) 'A model of currency depreciation and the debt-inflation spiral', *Journal of Economic Dynamics and Control*, 15: 151–77.

Obstfeld, M. and K. Rogoff (1996) *Foundations of International Macroeconomics*, Cambridge MA: The MIT Press.

Persson, T. and L. Svensson (1989) 'Why a stubborn conservative would run a deficit: policy with time-inconsistent preferences', *Quarterly Journal of Economics*, 104: 325–45.

Persson, T. and G. Tabellini (eds) (1994) *Monetary and Fiscal Policy*, vol. 2. Cambridge MA: The MIT Press.

Persson, T. and G. Tabellini (1999a) 'Political economics and macroeconomic policy', in vol. 1C, Taylor, J. and M. Woodford (eds) *Handbook of Macroeconomics*, Amsterdam: North-Holland.

Persson, T. and G. Tabellini (1999b) 'The size and scope of government', *European Economic Review*, 43: 699–735.

Rodrik, D. (1997) *Where Did All the Growth go? External Shocks, Social Conflict and Growth Collapses*, Kennedy School of Government, Cambridge MA: Harvard University.

Rogoff, K. (1990) 'Equilibrium political budget cycles', *American Economic Review*, 80: 21–36.

Rogoff, K. and A. Sibert (1988) 'Elections and macroeconomic policy cycles', *Review of Economic Studies*, 55: 1–16.

Sargent, T. (1987) *Dynamic Macroeconomic Theory*. Cambridge MA: Harvard University Press.

Stokey, N. and R. Lucas (1989) *Recursive Methods in Economic Dynamics*, Cambridge MA: Harvard University Press.

Tanzi, V. and L. Schuknecht (1997) 'Reconsidering the fiscal role of government: the international perspective', *American Economic Review*, 87: 164–72.

Tanzi, V. and L. Schuknecht (2000) *Public Spending in the Twentieth Century*, Cambridge: Cambridge University Press.

6 The importance of being unimportant

The political economy of trade and growth in small states

Harvey W. Armstrong and Robert Read

> Laws are generally found to be nets of such a texture as the little creep through, the great break through and the middle-sized alone are entangled in.
>
> William Shenstone, after Solon (*Oxford Dictionary of Quotations*)

The second half of the twentieth century saw a substantial rise in the number of small states in the global economy, as a result of both decolonisation and the disintegration of larger states. The strong growth performance of many of these small states appears to belie the economic orthodoxy of 'big is beautiful' and calls into question the critical role generally assigned to increasing returns to scale in determining the 'viable' economic size of a nation state. This growth success, achieved in spite of their small size, suggests that the analysis of the determinants of the economic performance of small states and improved understanding of their growth strategies may generate useful potential insights to other, larger, developing states.

This chapter reviews the theoretical and empirical literature relating to the determinants of growth in small states in the context of their salient characteristics and considers the political economy of their trade and growth strategies. The critical focus is therefore on the role of endogenous policies in this growth performance. The more successful small states have shown themselves to be adept at identifying and taking advantage of opportunities to engage in free-riding and rent-seeking activities in the global economy, made possible by virtue of their small size; 'the importance of being unimportant' (Demas, 1965). The first section outlines the orthodox economic approach to growth and its implicit assumptions concerning the size of states as well as advances in growth theory, specifically endogenous growth. The second section summarises the salient economic characteristics of small states and the implications for their growth including a review of empirical research on the structural determinants of growth in small states. The third section addresses the issues of rent-seeking and free-riding by small states in the international economy in the context of optimal endogenous policy determination and growth. The fourth section concludes.

Orthodox economic analysis of size and growth

Although almost half of the sovereign states in the global economy have populations below five million, the primary focus of orthodox models of economic growth and structural transition continues to be relatively large populous states. This section briefly outlines the orthodox economic approaches to smallness and structural transition as well as summarising recent theoretical developments in endogenous growth theory.

The theoretical paradigm of the 'small open economy' is frequently used in the economic analysis of international trade and macroeconomic policy issues. This approach has widespread applicability, primarily because very few countries – as opposed to trading blocs – actually possess sufficient international market power to affect their own terms of trade.

Growth and structural transition in economic development

Orthodox economic models of growth, development and structural transition are founded upon several critical assumptions: the presence of a relatively large population; a large backward (traditional) agricultural sector; and a large agricultural labour force. Growth involves structural transformation from primary sector dependence, principally agriculture, into manufacturing through industrialisation and later into services. Structural transition has its origins in early work on growth (Lewis, 1955; Rostow, 1960; Gerschenkron, 1962) and has subsequently been formalised, theoretically and empirically (e.g. Chenery and Taylor, 1968; Kuznets, 1971; Chenery and Syrquin, 1975; Chenery *et al.*, 1986). Economic growth and development in these models is viewed as the result of industrialisation, whereby surplus low-productivity unskilled and under-employed agricultural labour moves to the large-scale labour-intensive and more technology-intensive manufacturing sector. This has the effect of raising average productivity per worker and aggregate output without significantly reducing agricultural production (Lewis, 1955).

These orthodox economic models of structural transition cannot be easily applied to small states because of their small populations and the consequent constraints on the practicalities of large-scale labour-intensive manufacturing. This requires the consideration of alternative approaches to growth, development and structural transition for smaller states; noted by Lewis because of his own experience of the Caribbean, but generally ignored subsequently. This point provides the basis for the analysis of the growth implications for small states in the context of their economic sub-optimality.

New growth theory and endogenous growth models

The recent revival of interest in formal neo-classical economic growth owes much to the incorporation of additional structural and 'conditioning' variables into conditional convergence models of growth, such as Barro (1991) which make use of increasingly widespread comprehensive harmonised international data sets. The

models provide a means to analyse the impact of specific variables, including endogenously-determined policy variables, on relative rates of growth between countries and so formalise Gerschenkron's concept of 'catch-up' in growth.

The use of endogenous growth models imposes no structural pre-conditions upon the countries being analysed. Instead, they rely upon theoretical justification for the use of specific conditioning variables and the availability of appropriate statistical data. Endogenous growth models therefore offer a potential means to investigate the determinants of growth in small states. The endogenous growth literature identifies a number of important conditioning variables which affect a country's relative rate of growth (see Edwards, 1993; Temple, 1999), some of which are of critical importance with regard to small states.

Size constraints and growth implications for small states

The identification and analysis of the principal size constraints on the economic growth performance of small states is covered in an extensive literature on the subject (for example, Robinson, 1960; Demas, 1965; Benedict, 1967; Selwyn, 1975; *World Development*, 1980, 1993; Jalan, 1982; Dommen and Hein, 1985; Armstrong *et al*., 1996). This section provides a brief summary of the salient characteristics of small states in the context of their economic sub-optimality and considers the implications of small size for growth.

The critical size constraints on small economies

The sub-optimality approach, whereby small states are inherently constrained by insufficient internal economies of scale, provides an appropriate context for the identification of their salient economic characteristics and their policy limitations.

The small size of the domestic market

Domestic demand is insufficient to reach the minimum efficient scale of production such that the unit cost of a wide range of goods and services, if produced locally, is relatively high. While the impact of this constraint is uneven given different cost structures, it severely inhibits the evolution of large-scale manufacturing. Even where production is possible, the small domestic market limits the numbers of feasible firms, with adverse implications for competition and therefore prices. A small domestic market is also likely to impede indigenous R&D and technical progress (Briguglio, 1995).

The limited domestic resource base

Small size does not preclude extensive and valuable natural resource endowments, but it may limit the availability of the capital necessary for their exploitation. A more critical resource problem is the scarce domestic supply of labour which precludes standard structural transition as a development strategy. Small

states cannot feasibly specialise in the large-scale production of low-skill labour-intensive goods. Instead, their growth is expected to be heavily reliant upon human- and physical-capital-intensive activities (Bhaduri *et al*., 1982). Notwithstanding, scarce job opportunities in poorer small states has led to large migratory outflows and a growing reliance upon foreign remittances. More successful small states, however, have resorted to temporary in-migration to resolve shortages of particular specialist skills.

The narrow structure of output and exports

The small domestic market, limited resource base and the high cost of providing low volumes of a wide range of goods and services constrain the range of viable domestic productive activities. Further, increasing returns necessitates specialisation, leading to a relatively undiversified structure of output and export activities. This concentration may be compounded by geographic export concentration whereby the low absolute volume of exports means a reliance upon a limited number of export markets. A high degree of export concentration exposes an economy to unpredictable exogenous shocks via export price and earnings instability, exacerbated in the case of small states by their lack of local inter-industry linkages. Small states are also highly susceptible to import shocks because of their high degree of import dependence. The standard solution to these problems is increased self-sufficiency and diversification but small states have only limited potential to do so because of their size and resource constraints.

The high structural openness to trade

The narrow structure of domestic output means that there is a significant production-consumption asymmetry in small states since most domestic consumption is accounted for by imports (Kuznets, 1960). The critical structural importance of tradeables necessitates the pursuit of an extremely open trade regime and a high degree of integration with the international economy. Small states possess few, if any, sources of autonomous internal growth such that there is only limited scope for endogenous (functional) trade policy to promote import substitution and infant industry activities. Openness also has important implications for macroeconomic policy-making since any instability in import prices and/or export earnings rapidly feed through into the balance of payments and the net foreign exchange position. Since foreign currency transactions predominate, there is also a high degree of international monetisation in the economy which constrains domestic monetary autonomy because of thin currency markets and exchange rate and foreign reserve volatility (Ally, 1975; Helleiner, 1982). These problems can be avoided by the adoption of a more managed exchange rate policy via membership of a hard currency area which acts as a low-cost automatic stabiliser providing insulation against external volatility. In the absence of capital controls, this comes at the expense of monetary sovereignty including the determination of interest rates and inflation (Khatkhate and Short, 1980).

The implications of size for growth in small states

It is evident that small states experience critical constraints on the development of a critical mass of domestic economic activities because of their small size. This, in turn, has the effect of limiting the choice of growth strategy and the spectrum of policies with which they can be achieved. Small states therefore face significant growth challenges over and above those of larger countries.

The impact of the size constraints on the growth of small states can be conceptualised in terms of their greater degree of embodied risk, generally referred to as vulnerability. Economic vulnerability arises primarily because of the high structural openness and dependence upon trade of small states. This is often compounded by political, strategic, ecological and meteorological vulnerability, especially in the case of small island states because of their topography and remoteness (Commonwealth Consultative Group, 1985; Briguglio, 1995). The susceptibility of an economy to exogenous shocks of whatever kind increases the risk of short-run volatility around the underlying trend rate of growth. The extreme vulnerability of many small states therefore suggests that they are likely to experience lower long-run average rates of growth than larger states, characterised by greater amplitudes of fluctuation. In addition, many small developing states may be unable to afford precautionary insurance to fully compensate for the impact of any volatility because of scarce financial resources. Small states must therefore necessarily establish a trade-off between the increased risks of excessive domestic specialisation and the high costs of a diversified production base.

The general tenor of this discussion of the economic constraints on small states is that their size is disadvantageous, particularly with respect to their vulnerability. It is possible, however, to assert that some characteristics of small states may, in fact, be advantageous and actually promote growth. They lack the burden of a significant traditional backward agricultural sector characterised by subsistence output, low incomes, low productivity and a lack of technology. Their high degree of structural openness to trade also offers considerable potential growth benefits because of its domestic multiplier effects (Ashoff, 1989) although there remains some scope for endogenous trade policy. Many small states are also claimed to have a relatively high degree of internal social cohesion because of their size which encourages the creation of social capital. Given greater communal involvement and consensus decision-making, they therefore have the potential to be more responsive to change and more flexible in their policy-making, so providing a fertile environment for economic growth.

It is also useful to consider the inferences for growth in small states to be derived from endogenous growth theory. Three critical conditioning variables can be seen to be of particular importance. Openness to trade has beneficial effects on domestic competitiveness through imports which embody high levels of technology. A greater stock of human capital is argued to increase long-term productivity and growth because it is not subject to diminishing returns. Human capital formation has been identified as a key source of comparative advantage in small states such that investment in education, training and learning-by-doing can be expected

to have important long-run growth and productivity effects. Interaction between openness and human capital is of particular importance, given increasing returns in R&D and innovation (Briguglio, 1995), since the absorptive capacity for technology embodied in imports is dependent upon the endowment of human capital. Finally, the existence of 'convergence clubs' suggests that close regional proximity to prosperous larger economies may also benefit small states.

This discussion suggests that, in spite of the many disadvantages imposed upon small states by their size, they also enjoy some potential advantages. Further, the insights of endogenous growth theory in particular suggest that small states are well-placed to achieve high rates of growth in spite of their economic sub-optimality because of their high degree of openness to trade and propensity for human capital formation. While small states therefore face major growth challenges, they also have significant scope for optimal endogenous policy formulation to at least partially offset the adverse effects of their small size. Strategic behaviour by small states in the international economy may therefore be an important factor in explaining their growth success.

The growth performance of small states

It is evident that while small size inhibits the economic structure of small states, the strong growth performance of a significant sub-set of small states suggests the influence of successful countervailing growth strategies. There is abundant anecdotal evidence on the growth success of many small states, for example Luxembourg, Singapore and Mauritius, but only a very limited empirical literature investigating the underlying determinants of their economic performance. This is partly a consequence of severe data problems such that empirical analyses are dependent upon available data sets for a limited range of economic variables. Robust empirical analysis therefore requires appropriate statistical techniques and a need to avoid problems of sample selection bias.

Several studies examine the impact of size on growth, but no significantly negative relationship is found, irrespective of the comprehensiveness of their data sets and size definitions (Khalaf, 1971; Blazic-Metner and Hughes, 1982; Milner and Westaway, 1993; Armstrong *et al*., 1996, 1998). Given that small size imposes significant economic constraints on the growth of small states, the lack of a generally consistent cross-sectional pattern of economic performance according to size needs to be explained. One important explanation is that some small states perform extremely well in spite of their size while many larger states perform worse than expected. This argument is supported by the presence of unexpectedly large numbers of small states in the World Bank's highest income categories and disproportionately fewer in the lowest income categories (Armstrong *et al*., 1998). Further, the impact of size on growth may, in many of cases, be outweighed by the effects of other factors, such as the quality of endogenous policy-making.

It has been argued that small size may be neither necessary nor sufficient for low rates of growth and levels of income (Srinivasan, 1986) since trade enables small states to overcome their smallness by widening the extent of their market. An open

trade policy promotes comparative advantage through its effects upon domestic prices, efficiency and competitiveness. In small states, these effects are intensified because of the dominance of tradeables and their contestable impact upon domestic competition (Armstrong and Read, 1998a). The need for small states to adhere very closely to their fundamental comparative advantage suggests that they are more likely to benefit from the high growth effects of trade. The comparative advantage of small states can be expected to lie in some natural resources together with relatively human capital-intensive manufacturing and service activities not subject to significant scale economies. This view of the sectoral determinants of growth in small states is confirmed by statistical analysis using regional and global data sets. The findings indicate the strong contribution to growth in small states of a rich natural resource base together with a strong service sector, notably in both financial services and tourism, while agriculture has low growth effects (Armstrong and Read, 1995; Armstrong *et al*., 1996, 1998). These empirical findings are supported by the simple structural typology developed by UNCTAD (1997) and extended in Armstrong and Read (1998b). This typology suggests that specialisation in service activities in particular appears to have had a strong positive impact upon the economic performance of small states, that diversification has had stabilisation effects and that growth has generally been unhampered by exogenous shocks. These findings also suggest that, as expected, structural transition in small states differs fundamentally from the orthodox model of economic development (see Read, 2002a).

The literature on small states pays very little attention to the role of location, as opposed to isolation, in spite of the significance of 'convergence clubs' in endogenous growth models. Empirical investigations of the effect of broader regional location on growth, using World Bank regional definitions, find that it has a significant impact on income levels in small states (Armstrong *et al*., 1996, 1998; Armstrong and Read, 2001). This regional effect occurs in spite of the globalisation process and can be at least partially explained by the critical importance of trade to small states and the substantial benefits of being located close to prosperous and high growth markets.

Islandness is often regarded in the literature as a critical factor in the growth of small states. It is important to note, however, that its impact on growth is found to be only weakly negative (Armstrong *et al*., 1996, 1998).

Free riding and rent-seeking: small states and the international order

The profile and influence of small states in the international economic and political order has been growing in recent years, primarily because of the rise in their absolute numbers and increasing cohesion as a bloc to defend their mutual interests. Within the UN, the Association of Small Island States (AOSIS) co-ordinates the interests of many sovereign small member states. A special UNCTAD (UN Conference on Trade and Development) department has been created to look after the interests of small island developing states (SIDS). Small states also form an influential bloc within the Commonwealth which produces the only statistical

publication devoted solely to small states and has actively promoted extensive research on vulnerability.

The decision-making structure of many international institutions, notably the UN and its associated organisations, is founded upon the principle of one country – one vote, regardless of size. Members' budgetary contributions are generally both size- and means-related, subject to a minimum threshold contribution for the smallest states. Individual sovereign small states therefore possess the potential block voting power to 'punch above their weight' in these organisations. The growing participation of small states in the major international organisations is leading to a gradual shift in the balance of power within these institutions as their collective power has risen. Small states are therefore increasingly able to exercise discernible influence over the decision-making process and the shaping of the international regulatory framework in spite of their small aggregate budget contributions. Concern at this trend was first expressed more than three decades ago (UNITAR, 1969). The potential for favourable policy bias suggests that there may be considerable rent-seeking opportunities for small states to be derived from the policy stance of these institutions.

Rent-seeking behaviour of this type has recently come to the fore regarding the use of vulnerability indices, the choice of variables and their respective weightings (Briguglio, 1998). There has been some conflict, both between small states and between small states and larger developing countries, over the magnitude and impact of vulnerability and therefore their respective eligibility for international financial assistance. The pursuit of self-interest in such circumstances may be counter-productive however, to the extent that any divisions may disrupt the cohesion of other effective pressure groups involving these states.

Free-riding: defence as an international public good

The strategic vulnerability of small states together with their dependence upon international trade suggests that they are likely to be strong advocates of a peaceful global environment. Most small states however, lack the resources and a sufficient critical mass to provide effective defence measures. Although defence is a national public good, it requires a disproportionate share of domestic resources (Kuznets, 1960), given the critical importance of absolute size and the presence of increasing returns, and might still not be an effective deterrent. Many small states therefore prefer to eschew costly and excessively onerous domestic defence expenditure altogether. Instead, they rely upon the commitment of the UN (or its larger members) and strategic alliances with larger states to uphold their national sovereignty and territorial integrity.

World peace is an international public good which benefits all (or most) countries but cannot be maintained without the altruistic commitment of a core group of countries to maintain international peace, and act as global law enforcers thus defending smaller and weaker states. The funding and provision of defence goods therefore tends to be asymmetric between countries. This generates significant positive externalities for those countries, such as small states, for whom the cost of effective deterrence is beyond their means in the form of a defence 'umbrella'.

Given their small size, the lack of any substantial domestic defence commitment by most small states can therefore be viewed as a logical policy-decision with significant potential financial benefits in terms of a reduced domestic tax burden. This means that a greater proportion of government expenditure can be directed towards social public goods such as education and health – this may contribute to the high human development index scores of many small states. Free-riding on the international defence community, however, has its own associated costs because of the strategic vulnerability of small states; the failure of international deterrence may result in the temporary or even permanent loss of sovereignty, conflict and the loss of life. The position of Kuwait in the 1991 Gulf War is a case in point.

Free riding: global trade liberalisation as an international public good

Small states can be expected to be the greatest beneficiaries proportionately from a peaceful global environment because it facilitates a relatively liberal international trade order. Global trade liberalisation tends to favour small states because of their high degree of structural openness even though they have little actual influence over the liberalisation process itself. The GATT/WTO articles on the conduct of international trade incorporate the principles of reciprocity and multilateralism, the combined effect of which was specifically designed to generate positive externalities for smaller and weaker countries through the process of trade liberalisation. Reciprocal concessions made as a result of negotiations between the major industrialised countries, the EU, Japan and the US in particular, apply multilaterally to all other member states. Countries with little bargaining power therefore benefit from significant multilateral trade concessions which they would otherwise have been unable to wrest, individually or collectively. Small states are therefore able to free-ride on the positive externalities explicitly designed to be generated by the global trade liberalisation process.

These externalities, however, are not necessarily always positive. A key objective of international trade reform under the WTO, including GATS (General Agreement on Trade in Services), is the creation of a 'level playing-field' to promote comparative advantage and competition. Many small states rely upon bilateral, as well as multilateral, trade agreements with their major trading partners, and are likely to be less able to compete head-on with many larger countries in a more openly competitive liberal global trading environment (Armstrong and Read, 1998b). The drive towards multilateralism together with the phasing-out of bilateral agreements as incompatible with most favoured nation (MFN) treatment is therefore likely to increase the competitive pressure on small states as some niche markets are lost. Further, many of the undertakings associated with trade liberalisation, such as the GATS, TRIMs (Trade-Related Investment Measures) and TRIPs (Trade-Related Intellectual Property Rights), require substantial administrative and technical commitments with significant resource implications. Many of these legal commitments, however, are irrelevant to the structures and needs of small states, particularly small developing states, and represent a heavy burden on limited domestic resources. Although these shortcomings reduce the magnitude of

the positive externalities for small states from the liberal international trade order, the alternative option of remaining outside may be very costly.

Free-riding: pseudo-monetary integration and currency unions

Small size has important implications for domestic exchange rate and monetary policy in small states. This is because structural openness and the significance of imports in domestic consumption raises the domestic multiplier effects of exogenous shocks. Under a flexible exchange rate regime, there is limited scope for adjustment in a small state since the exchange rate is effectively determined exogenously. This is because it is difficult to maintain internal and external balance simultaneously in the absence of capital controls (Helleiner, 1982) and is exacerbated by thin currency markets and a dependence upon a limited number of trade partners. Small states with their own independent currencies will therefore tend to be affected by severe structural problems and exchange rate volatility, both of which feed through into internal cost and price instability effects (Chadha, this volume).

For these reasons, most small states choose to adopt a more 'managed' exchange rate policy, whether through formal or informal links with hard currencies. A significant number have ceded currency autonomy in favour of pseudo-membership of a hard currency area, most commonly with the US dollar (Read, 1995; Armstrong and Read, 1998b) – Ecuador's recent adoption of the dollar demonstrates that this is not the exclusive preserve of small states. The monetary integration is 'pseudo-' because membership is not federal and monetary autonomy remains with the national central bank of the hard currency. Pseudo-membership of a currency union provides an automatic stabiliser which insulates members' economies against external volatility (excluding that of the chosen hard currency). The loss of the exchange rate as a policy instrument means the loss of significant endogenous policy autonomy with respect to the determination of the domestic rate of interest and inflation and, in the absence of capital controls, over output and aggregate demand given that the relative price of tradeables and non-tradeables is fixed (see Khatkhate and Short, 1980). This imposes substantial fiscal rectitude, reflected in reduced external indebtedness and deficit financing. In effect, pseudo-monetary integration therefore enables small states to free-ride on the credibility of a third country's currency. Such free-riding is beneficial to the third country to the extent that it stimulates greater international demand for its currency (seigniorage gains) so long as its own long-term macroeconomic policy objectives are not destabilised.

Many other small states pursue some form of fixed exchange rate policy with respect to one or more hard currencies, generally at a par value or via a long-term fixed exchange rate. In the slightly different cases of the CFA franc and the East Caribbean dollar, groups of small (and larger) states are in currency union with a regional currency that is itself fixed to a hard currency (the French franc and US dollar respectively). The advent of the Euro in 2002 will therefore bring the exchange rates of many small states into some form of fixed alignment with the EU and

with each other. The post-Bretton Woods flexible exchange rate regime between the leading industrialised nations has therefore had unforeseen secondary effects. The growing importance of regional currency blocs, whether pseudo-unions or not, means that fluctuations between the major currencies have widespread ramifications for many third countries.

The success of the offshore financial service sector in many more prosperous small states is partly based upon pseudo-integration with a hard currency. The sector itself therefore free-rides on the positive externalities generated by the credibility, commitment and macroeconomic policy of the third country's central bank. This strategy means that the offshore financial services sector in a small state can derive greater effective fiscal probity and significantly reduce risks to investors of both currency volatility and rates of return.

Rent-seeking: the bilateral trade relations of small states

International trade is critical in fostering growth in small states, but exposes their economies to inherent risks of exogenous shock. This vulnerability can be partially offset by assuring access for their exports in the markets of their principal trading partners. Dependence upon a limited number of export markets, however, also increases the exposure of small states to the risk of exogenous shocks in these markets. This suggests that small states should initiate policy strategies designed to deepen the extent of their integration with their major trading partners, including trade agreements to assure preferential market access as well as pseudo-monetary integration to dampen the effects of any macroeconomic instability.

Many small states have been highly successful in securing asymmetric (non-reciprocal) bilateral trade concession arrangements with their larger neighbours, including regional trade blocs. This is somewhat contrary to expectations given the relatively weak bargaining power of small states with respect to their larger trading partners. One explanation for the proliferation of such preferential bilateral arrangements is their relative unimportance. Although the value of such bilateral arrangements to small states may be significant in that they provide more assured market access and reduced risks associated with niche export strategies, the cost to the donors may be negligible. The EU for example, has granted many such bilateral concessions to non-member small states in Western Europe (Armstrong and Read, 1995), generally in the form of non-reciprocal free access to the EU market and derogations for some sensitive exports.

Bilateral trade arrangements of this type, however, are incompatible with the international regulatory framework for trade established under the GATT and enforced by the WTO. They are therefore highly susceptible to the ongoing process of trade liberalisation, particularly given the WTO's more effective dispute settlement procedures. Strict adherence to axiomatic multilateralism and the creation of a level playing-field for international trade is therefore likely to deprive small states of many niche opportunities by removing marginal but critically important sources of protection which contribute to their growth success.

Rent-seeking: small states and economic integration

The discussion of pseudo-monetary integration suggests that small states might derive similar benefits from participating in regional integration schemes. Membership of free trade areas and trade blocs offers greater security of market access than bilateral arrangements and is a means by which vulnerability can be reduced through closer trade links and greater economic interdependence with neighbouring states. Orthodox economic theory generally argues that small states are the greatest beneficiaries from regional integration schemes because of the increased extent of their domestic market, so counteracting the adverse scale effects of small size, and their improved terms of trade (Graham, 1923; Balassa, 1967).

Small states might, on this basis, therefore be expected to be keen participants in regional integration schemes but this is not in fact the case. Luxembourg is currently the only small state full member of the EU (see Armstrong and Read, 1995 on these issues). Outside Western Europe, only Singapore and Brunei in APEC, the anglophone Caribbean states in CARICOM, and Botswana, Lesotho and Swaziland in SADEC are small state members of functioning regional trade blocs. Singapore was part of the Malay Federation at independence but quickly broke away. The West Indian Federation, created out of the larger British Caribbean colonies, collapsed into separate states soon after independence. The Free Trade Area of the Americas (FTAA) will be the first time that a significant number of small states have been involved in a regional integration scheme (Read, 2002b).

The dearth of small member states of regional integration schemes suggests that the apparent gains from membership are not necessarily clear-cut. Economic integration requires trade policy autonomy to be relinquished, a critical economic policy variable in small states. The view that small states are likely to be the greatest beneficiaries from economic integration is predicated on pre-integration trade regimes being highly protectionist such that there are significant gains from intra-regional trade liberalisation. Small states, however, necessarily exhibit a high degree of openness. Any gains may therefore be relatively small while adverse trade diversion effects could be very large because the discriminatory external policy stance of a trade bloc may reduce the extent of the market of small states (Rothschild, 1944, 1963). Economic integration may therefore reduce the welfare of small state members because of their critical loss of trade policy autonomy, particularly with respect to control over import-sourcing and negotiating bilateral agreements with key trade partners (Read, 1995, 2002b). Further, the agglomerative impact of inter-regional growth effects may favour larger more established industrialised regions at the expense of peripheral regions (Rothschild, 1944), particularly if such regions lack appropriate policy autonomy (Hirschmann, 1958).

Small states, particularly relatively successful ones, are therefore unlikely candidates for regional integration schemes because of the primacy of their trade policy autonomy. Bilateral trade arrangements would therefore appear to be a more rewarding strategy since they allow them to retain this policy autonomy (Rothschild, 1963). Nevertheless, for less successful small states, there is significant scope for the balance of the costs and benefits of integration to vary considerably for a num-

ber of reasons: a relatively protectionist trade policy stance prior to integration; the prospect of high regional growth effects; the potential for significant fiscal transfers (EU structural funds); and domestic political reasons (e.g. Cyprus).

Rent seeking: differences between national regulatory frameworks

Sovereignty over domestic policy-making enables jurisdictions to establish distinct national regulatory frameworks for a wide range of economic activities. Many of the more successful small states specialise in niche offshore service activities, particularly in the areas of finance, insurance and ship registration, which generate high revenue and growth effects. Their international competitiveness in these sectors is at least partly dependent upon strategic taxation regimes and differentiated regulatory frameworks. The attraction of offshore financial centres in many small states is further enhanced by the local use of hard currencies which reduce risks and increase returns.

Tax competition with onshore jurisdictions generates significant marginal revenue gains for small states based upon their comparative advantage in human capital and fiscal policy autonomy. Recent efforts by the EU and OECD to outlaw 'unfair' tax competition from tax havens have been partly motivated by the need to eradicate international money laundering but also by the loss of domestic (onshore) tax revenue. It is not clear what is meant by unfair in this context. Many small states have demonstrated their fiscal probity by meeting the rigorous requirements of the EU's Financial Services Directive and there remains a suspicion that the leading industrialised countries dislike competition for investment funds. Illegal financial activities and enforcement difficulties are present in weaker jurisdictions, including some small states. Nevertheless, the success of many offshore financial services sectors in small states provides no adequate reason for international tax harmonisation. At the same time, the industrialised countries appear unwilling to extend the level playing-field in trade in goods to financial services in spite of the GATS (Read, 2001).

Rent-seeking: the internationalisation of labour markets in small states

The critical constraint on labour has important implications for the structure of economic activity and comparative advantage in small states. There is an additional dimension to this discussion, however, in that their labour markets exhibit a high degree of internationalisation in terms of migratory flows. These flows occur in both directions, with sizeable out-migration from poorer small states and sizeable in-migration to wealthier ones.

Out-migration, usually of relatively unskilled labour, is often the result of the dearth of domestic economic activity and a consequent lack of employment opportunities in poorly-developed small states, notably the Pacific islands. Migratory outflows are greatly restricted by immigration policies in many industrialised destination countries, but former colonial links and the zones of free labour mobility

provide limited opportunities – such as between some Pacific islands and Australia, New Zealand and the US. Significant out-migration has led to some small states becoming almost completely dependent upon the contribution of remittances from abroad to national income – migration, remittances and barter (MIRAB) economies (Type I economies in the UNCTAD typology).

In-migration to small states is facilitated by their domestic policy autonomy over immigration controls. Many relatively wealthy small states operate highly selective immigration policies which enable them to relieve skills deficits and domestic shortages of unskilled labour through the judicious use of work permits, temporary residency and other policy instruments, for example, Andorra and Liechtenstein. Policy autonomy over labour inflows is also exercised by many non-sovereign small states in Western Europe, notably the Channel Islands, Gibraltar and the Isle of Man.

While the regulation of international labour migration is generally characterised by a significant asymmetry between out-migration and in-migration, many small states have shown themselves to be adept at capitalising on their own relative unimportance. Outflows of labour have been aided by traditional links and bilateral agreements with larger industrialised countries as well as intra-regional mobility, such as inter-island migratory labour flows in the Caribbean. Out-migration has therefore provided a means of reducing unemployment and raising national income in poorer small states. Policy autonomy over in-migration has been a very effective tool for relieving gaps in the domestic labour markets of small states.

Rent-seeking: location-specific niches

The strategic geographic location of many small states, particularly islands, was originally a prime source for generating income, encompassing logistics as well as military objectives and communications. Advances in long-distance communications have reduced this role but some small states use their strategic location for rent-seeking through links with their former colonial powers, ideological alliances and/or by playing-off the major powers against each other. Unique environmental resources and the growth of long-haul tourism, however, are increasingly important sources of location-specific income in remote islands (Wace, 1980). Many small states also make use of their autonomous status to generate income from the issue and sale of special series of coins and stamps for collectors (Connell, 1988).

Rent-seeking: trends in sub-national regional policy autonomy

The preceding discussion of international political economy is primarily concerned with the policy strategies of small jurisdictions. It also has wider ramifications in the context of regional devolution within larger states such that the concept of policy sovereignty is becoming increasingly blurred. At the same time, there is increasing divergence between the political and economic functions of states. Political autonomy is being increasingly devolved to relatively small and distinct sub-national groups simultaneous with the growing participation of sovereign states in regional

economic blocs as global trade liberalisation progresses (Naisbitt, 1994; Svetlecic, 1996). These developments are complementary to the general discussion of policy autonomy in small states in that they highlight the selective retention of policy autonomy combined with the ceding of sovereignty in less crucial spheres.

Conclusions

This chapter considers the international political economy of small states in the context of the structural and policy implications of their salient economic characteristics. In particular, it highlights the critical role of policy autonomy in influencing their growth performance through international free-riding and rent-seeking activities. International political economy is an integral component in the optimal policy portfolio of small states and thus a key element in their growth strategy, implemented through a combination of policy autonomy and the ability to pursue niche activities.

Many small states have successfully attained sustained economic growth and high levels of per capita income in spite of the constraints of their size and vulnerability. Their long-term growth is by no means assured, however, since many of the principal sources of growth in small states are subject to uncertainty and change. This includes the trend towards multilateralism at the expense of bilateralism and attempts by the EU and OECD to eliminate harmful tax competition. These developments may force small states to seek shelter in large regional trade blocs in spite of their evident shortcomings.

The move towards creating a level playing-field in the international economy can therefore be expected to place many small states at a disadvantage relative to most larger states because of their small size, insufficient economies of scale in many activities and, in many cases, their relative isolation/remoteness from major markets. It remains to be seen whether individual small states can build upon their internal cohesion and flexibility to respond to these challenges and maintain their niche positions in the international economy. This may depend upon their success in co-ordinating and influencing international policy decisions. Otherwise, they may return to being marginalised by the globalisation process and condemned to economic peripheralisation and stagnation.

Acknowledgements

The authors are grateful for comments and insights from participants at the ISISA V Islands Conference, Mauritius, 1–5 July 1998 and the UK IESG Easter Conference, University of Warwick, 15–16 April 1999. Particular thanks are due to Godfrey Baldacchino, Barry Bartmann, Lino Briguglio, David Milne and Mansoob Murshed.

References

Ally, A. (1975) 'The potential for autonomous monetary policy in small developing countries', in P. Selwyn (ed.) *Development Policy in Small Countries*, Beckenham: Croom Helm.

Armstrong, H.W., R. Jouan de Kervenoael, X. Li and R. Read (1996) *The Economic Performance of Micro-States*, London: Report for the UK ODA.

Armstrong, H.W., R. Jouan de Kervenoael, X. Li and R. Read (1998) 'A comparison of the economic performance of different micro-states and between micro-states and larger countries', *World Development*, 26 (4), 539–56.

Armstrong, H.W. and R. Read (1995) 'Western European micro-states and EU autonomous regions: the advantages of size and sovereignty', *World Development*, 23 (8): 1229–45.

Armstrong, H.W. and R. Read (1998a) 'Trade, competition and market structure in small states: the role of contestability', *Bank of Valletta Review*, no. 18 (December): 1–18.

Armstrong, H.W. and R. Read (1998b) 'Trade and growth in small states: the impact of global trade liberalisation', *World Economy*, 21 (4): 563–85.

Armstrong, H.W. and R. Read (2001) 'Explaining differences in the economic performance of micro-states in Africa and Asia', in P. Lawrence and C. Thirtle (eds) *Africa and Asia in Comparative Development*, Basingstoke: Palgrave.

Ashoff, G. (1989) *Economic and Industrial Development Options for Small Third World Countries*, Occasional Paper no. 91, Berlin: German Development Institute.

Balassa, B. (1967) *Trade Liberalisation Among Industrialised Countries: Objectives and Alternatives*, London: McGraw-Hill.

Barro, R.J. (1991) 'Economic growth across a cross-section of countries', *Quarterly Journal of Economics*, 85: 407–33.

Benedict, B. (ed.) (1967) *Problems of Smaller Territories*, London: Athlone Press.

Bhaduri, A., A. Mukherji and R. Sengupta (1982) 'Problems of long-term growth in small economies: a theoretical analysis', in B. Jalan (ed.) *Problems and Policies in Small Economies*, Beckenham: Croom Helm for the Commonwealth Secretariat.

Blazic-Metner, B. and H. Hughes (1982) 'Growth experience of small economies', in B. Jalan (ed.) *Problems and Policies in Small Economies*, Beckenham: Croom Helm for the Commonwealth Secretariat.

Briguglio, L. (1995) 'Small island developing states and their economic vulnerabilities', *World Development*, 23 (10): 1615–32.

Briguglio, L. (1998) 'Towards the construction of an economic vulnerability index: recent literature with special reference to small island developing states', paper presented at the UK International Economics Study Group (IESG) Conference, Small States in the International Economy, University of Birmingham, 16–17 April.

Chadha, J (2002), 'Some observations on small-state choice of exchange rate regime' this volume.

Chenery, H.B., S. Robinson and M. Syrquin (1986) *Industrialisation and Growth: A Comparative Study*, Oxford: Oxford University Press.

Chenery, H.B. and M. Syrquin (1975) *Patterns of Development: 1950–1970*, Oxford: Oxford University Press.

Chenery, H.B. and L. Taylor (1968) 'Development patterns: among countries and over time', *Review of Economics and Statistics*, 50 (4): 391–416.

Commonwealth Consultative Group (1985) *Vulnerability: Small States in the Global Society*, London: Commonwealth Secretariat.

Connell, J. (1988) 'Sovereignty and survival: island micro-states and the third world', Monograph No. 3, Sydney: University of Sydney, Department of Geography.

Demas, W.G. (1965) *The Economics of Development in Small Countries: With Special Reference to the Caribbean*, Montreal: McGill University Press.

Dommen, E.C. and P.L. Hein (eds) (1985) *States, Microstates and Islands*, London: Croom Helm.

Edwards, S. (1993) 'Openness, trade liberalisation and growth in developing countries', *Journal of Economic Literature*, 31 (3): 1358–93.
Gerschenkron, A. (1962) *Economic Backwardness in Historical Perspective*, Cambridge MA: Harvard University Press.
Graham, F. (1923) 'The theory of international values re-examined', *Quarterly Journal of Economics*, 37: 54–86, reprinted in H. Ellis and L.S. Meltzer (eds) (1961) *Readings in the Theory of International Economics*.
Helleiner, G.K. (1982) 'Balance of payments problems and macro-economic policy', in B. Jalan (ed.) *Problems and Policies in Small Economies*, Beckenham: Croom Helm for the Commonwealth Secretariat.
Hirschmann, A.O. (1958) *The Strategy of Economic Development*, New Haven CT: Yale University Press.
Jalan, B. (ed.) (1982) *Problems and Policies in Small Economies*, Beckenham: Croom Helm for the Commonwealth Secretariat.
Khalaf, N.G. (1971) *Economic Implications of the Size of Nations: With Special Reference to Lebanon*, Leiden: E.J. Brill.
Khatkhate, D.R. and B.K. Short (1980) 'Monetary and central banking problems of mini-states', *World Development*, 8 (12): 1017–26.
Kuznets, S. (1960) 'The economic growth of small nations', in E.A.G. Robinson (ed.) *The Economic Consequences of the Size of Nations*, London: Macmillan.
Kuznets, S. (1971) *Economic Growth of Nations: Total Output and Production Structure*, Cambridge MA: Belknap.
Lewis, W.A. (1955) *The Theory of Growth and Development*, London: Allen & Unwin.
Milner, C.R. and T. Westaway (1993) 'Country size and the medium-term growth process: some cross-country evidence', *World Development*, 21 (2): 203–12.
Naisbitt, J. (1994) *Global Paradox*, London: Nicholas Bealey.
Read, R (1995) 'Micro-states and currency systems: a review of the issues and evidence', paper presented at the UK IESG Easter Conference, University of Manchester, 21 April.
Read, R. (2001) 'Small states and the proposals for international tax reform', mimeo, unpublished manuscript.
Read, R. (2002a) 'Growth, economic development and structural transition in small vulnerable states', in S.M. Murshed (ed.) *Globalisation, Marginalisation and Development*, London: Routledge.
Read, R. (2002b) 'Small states and the political economy of international integration: the Caribbean states and the Free Trade Area of the Americas', in N. Karagiannis and M. Witter (eds) *The Caribbean the Free Trade Area of the Americas and the Reality*, forthcoming.
Robinson, E.A.G. (ed.) (1960) *The Economic Consequences of the Size of Nations*, London: Macmillan.
Rostow, W.W. (1960) *Stages of Economic Growth*, Cambridge: Cambridge University Press.
Rothschild, K.W. (1944) 'The small nation and world trade', *Economic Journal*, 54: 26–40.
Rothschild, K.W. (1963) 'Kleinstaat und integration', *Weltwirtschaftliches Archiv*, 90: 239–75.
Selwyn, P. (ed.) (1975) *Development Policy in Small Countries*, Beckenham: Croom Helm.
Srinivasan, T.N. (1986) 'The costs and benefits of being a small, remote, land-locked or mini-state economy', *World Bank Research Observer*, 1 (2): 205–18.
Svetlecic, M. (1996) 'Global economic integration and political disintegration', in M. Svetlecic and H.W. Singer (eds) *The World Economy: Challenges of Globalisation and Regionalisation*, London: Macmillan.
Temple, J. (1999) 'The new growth evidence', *Journal of Economic Literature*, 37 (1): 112–56.
UNCTAD (1997) *The Vulnerability of Small Island Developing States in the Context of Globalization: Common Issues and Remedies*, Geneva: UNCTAD, SIDS.

UNITAR (1969) *Status and Problems of Very Small States and Territories*, New York: UN, republished as *Small States and Territories: Status and Problems*, New York: Arno Press (1971).

Wace, N. (1980) 'Exploitation of some of the advantages of remoteness and isolation in the economic development of Pacific islands', in R.T. Shand (ed.) *The Island States of the Pacific and Indian Oceans: Anatomy of Development*, Development Studies Centre Monograph no. 23, Canberra: Australian National University.

World Development (1980), vol. 8 (12), special issue.

World Development (1993), vol. 21 (2), special issue.

7 Some observations on small state choice of exchange rate regime

Jagjit S. Chadha

> The great nations have always acted like gangsters, and the small nations like prostitutes.
>
> From the film Dr Strangelove, Directed by Stanley Kubrick (1963)

> Evidence that politics determine currency questions has generally been sought from recent or prospective examples of changes in existing circumstances . . . How many argue that certain small nations using the dollar, such as Liberia and Panama, belong optimally to the US currency area? Iceland (has) a separate currency for about 250,000 people, and Liechtenstein, San Marino and Monaco whose economies are so closely joined to their neighbours that having separate currencies would be economically inefficient. Size alone does not seem to be the most important criterion.
>
> Charles Goodhart (1995)

The question of whether to maintain a floating exchange rate, adopt some form of fixity or even some other state's currency is complex and, arguably, is not one on which we have any clear prescription. Even the eleven founder members of the European Monetary Union were generally regarded as embarking on a great adventure rather than any necessary economic policy. The theory of optimal currency areas, developed by Mundell (1961), has provided the starting point for much of the economic debate. We can sum up the resulting macroeconomic prescription as follows. Given sticky prices and imperfect factor mobility, the presence of significant real asymmetric shocks will lead to welfare gains when a floating exchange rate is maintained, against the alternative of a fixed exchange rate.[1] Against any potential gain from pursuing optimal macroeconomic stabilisation, we need to place losses related to microeconomic inefficiencies. Specifically, those resulting from the inefficient allocation of traded goods across countries in which relative prices evolve, as a result of exchange rate fluctuations, with greater uncertainty.

But it is also clear, from at least as early as Kenen's (1969) eclectic view, that the calculus of economic stabilisation and microeconomic efficiency do not

constitute the whole picture faced by policy-makers. This is possibly because the choice of exchange rate regime continues to be closely connected with the kind of political choices, suggested by both Kubrick and Goodhart, which economists have only recently started to re-examine. Goodhart implies that standard criteria are often inapplicable to small states. Small states rely on their relations with the outside world for economic solvency, and this affects every aspect of state decision-making. For example, reliance on external – both economic and political – relations may help to explain the degree of structural openness found in these economies. Alesina, Perotti and Spolaore (1995) and Alesina and Spolaore (1997) give an example of this recent strategic literature. They develop a promising theory on the equilibrium number of states, depending on the need to share in the provision of public goods, which may *inter alia* have implications for the theory of currency areas. It then seems likely that political choices may weigh quite heavily in the economic calculus of policy-makers in smaller states. Many post-war monetary arrangements for LDCs seemed closely related to political as much as economic links: for example, those of the East African Currency Union to the UK or the CFA franc zone to France.

How do the issues surrounding the choice of exchange rate regime account for the current set of choices made by countries? Though we should perhaps recall that a longer view suggests that economies seem to practice alternating cycles between periods of floating and fixed exchange rates – e.g. the inter-war years dirty float and post-1972 float versus the classical gold standard and Bretton Woods regime.[2] We should, therefore, be careful to draw conclusions from any single snapshot. But superficially, at least, floating exchange rates appear to be in the ascendancy at the moment.

The shift in macroeconomic policies in the 1980s and 1990s towards current and capital account liberalisation has perhaps tended to increase the incentives of maintaining *orthodox* macroeconomic policies and, though not necessarily causal, has been associated with a trend towards greater flexibility in exchange rate arrangements. Certainly the fixed exchange rate regimes of the immediate post-war period seemed linked to a plethora of domestic and international controls on transactions. Two snapshots are illustrative. By end-1989 around 70 per cent of the 153 countries surveyed by the IMF had fixed or managed exchange rates. But by end-1996, that fraction had fallen to 37 per cent of the 185 countries surveyed. Developing or transition economies are not outliers in this trend: at end-1975 developing countries with flexible exchange rate regimes accounted for less than 10 per cent of developing country trade and real GDP, by the mid-1980s these proportions were both around 40 per cent and by end-1996 the proportions were 65 per cent and 80 per cent respectively.[3]

In this chapter, we will therefore examine the choice of nominal exchange rates for small states and, in particular, aim to gather some stylised facts on the time series properties of nominal exchange rates, real exchange rates and inflation rates for these economies. Essentially we find little evidence of price stickiness in small states, certainly, in comparison to the nominal exchange rate. This means that such states have little to gain in terms of stabilisation policy by the employment of

independent monetary policy. Additionally, for them there is little to be gained in terms of credibility by a fixed exchange rate and we must look outside this type of theory to explain the high incidence of fixed exchange rate regimes.[4] The first section outlines a theoretical model on the choice of exchange rate. The second section examines the recent evidence on real exchange rate variability and sticky prices for these small states. The third section concludes by making some tentative observations on the policy choices facing these states.

Theory

In this section we shall examine the choice of exchange rate regime in terms of the credibility-flexibility trade-off. A standard open-economy model, based on Devarajan and Rodrik (1991), will be used to gain an insight into the parameters relevant for a small state when considering whether to fix or float its exchange rate. Equation (7.1) expresses the policy-maker's objective function in terms of (symmetric) losses from target inflation and output growth rates:

$$W = -\left\{(\pi - \pi^*)^2 + \phi(y - y^*)^2\right\} \tag{7.1}$$

where W is welfare, π is inflation, y is output growth, ϕ is the weight attached by the authorities on the real (output) relative to the nominal (inflation) target and $*$ represent target values. The growth rate is given by an IS equation of the form taken by (7.2):

$$y = \bar{y} + \alpha(r) + \beta(\varepsilon_t) \tag{7.2}$$

where $\bar{y}$ is the economy's steady-state growth rate, r is the real exchange rate ($r = e - p$, described below) and ε_t is a random variable with variance σ^2 representing a positive real shock such as a productivity or terms of trade shock, α and β are elasticity parameters. Prices denoted by p, include wages, are set per period by rational forward-looking agents in expectation of the level of the nominal exchange rate and of the real shock, according to:

$$p = E(e) + \omega E(\varepsilon_t) \tag{7.3}$$

where E is the expectations operator, e is the level of the nominal exchange rate and ω is the elasticity of domestic prices in response to a real shock. There are two consequences of e being set after p, the first term on the right-hand side of (7.3) drops out in expectation and secondly, policy-makers are able to set output growth (in the short-run) by creating a real exchange rate shock. In a sense, the setting of p for two periods is motivated by the view that prices are sticky and thus allow the authorities to set output in the short-run. Finally, equation (7.4) writes inflation as the weighted average of the increases in the price of home and (homogenous) overseas goods, where μ is the weight of home goods in the domestic price index and operates as an inverse proxy for openness in this set-up:

$$\pi = \mu p + (1 - \mu)e. \tag{7.4}$$

If we solve this model for the quadratic losses under fixed (f) and floating (nf) exchange rates we find the following expression for the net gain from fixed exchange rates:

$$W_f - W_{nf} = \alpha^2\phi^2 \left[\left\{ \frac{y^* - \bar{y}}{(1-\mu)} \right\}^2 - \left\{ \frac{\beta^2\sigma^2}{[\alpha^2\phi + (1-\mu)^2]} \right\} \right]. \tag{7.5}$$

The first term in the large square brackets of equation (7.5) represents the benefits from fixed exchange rates and the second term the costs. Equation (7.5) states that first, increases in α, β and σ^2 will tend to make it more likely that a floating exchange rate will be preferred. Second, the smaller the difference between $(y^* - \bar{y})$ the more likely it is that floating exchange rates will be preferred. This is because the smaller the perceived temptation for policy-makers to inflate the smaller the gain in reducing inflation losses from fixing the exchange rates. Third, for plausible parameter values, reductions in μ (greater openness) make it more likely that a floating exchange rate will be preferred.[5] Fourth, as regards ϕ: (i) if fixed exchange rates are preferred (i.e. $W_f - W_{nf} > 0$) then increases in ϕ will tend to make a fixed exchange rate more beneficial (because the perceived inflation bias would be higher); and (ii) if floating exchange rates are preferred, increases in ϕ are ambiguous; but (iii) because the derivative of (7.5) with respect to ϕ is unambiguously positive there must be a value of ϕ which makes the monetary authorities indifferent between fixed and floating rates.[6]

This type of model suggests that an open economy which is susceptible to large (asymmetric) real shocks and whose monetary authorities have a relatively low preference for output fluctuation, compared to inflation stabilisation, may tend to prefer floating exchange rates. It may be reasonable to suppose that such characteristics describe many medium and large economies and hence explains the *trend* towards floating exchange rates. It may also allow us also to suggest that states with fixed exchange rates are likely to be those with a perceived inflation bias (high ϕ) and/or where there is little output gain from nominal exchange rate flexibility.

It would be useful to apply the predictions of this model to small states and examine whether we can explain the choice of regime. The IS analysis is ambiguous because although σ^2 or real shocks are likely to be high for small states, the fact that these economies are price takers is likely to reduce the relevant output elasticities α and β. As pointed out by Romer (1993), openness and the inflation bias term are likely to be related so that there will be less of temptation to increase output and less of an advantage from exchange rate fixity. Each of these parameters could be measured empirically but recall that the building block of the model is that the authorities have some control over output by altering the nominal exchange rate, which leads to changes in the real exchange rate because prices cannot move in the same period. And if we find empirically that prices are not relatively sticky then the potential gains *prima facie* seem limited.

Another important issue follows. If prices are sticky the move from fixed to floating exchange rates will actually *cause* high real exchange rate variability as shocks through relative money supplies are not automatically accommodated. On

the other hand if prices are not sticky then the persistent variability of the real exchange rate is likely to result from real innovations to the real exchange rate.

The view of flexible exchange rates in a world of sticky goods market prices combining to generate variable exchange rates can be tested by examining the relative contribution of nominal effective exchange rates and relative inflation rates towards real exchange rate variability, and is the purpose of the next section.

Empirical results

Summary statistics

According to the *World Tables* there are 50 states with populations of approximately 3.5 million or less. They seem to be reasonably distributed across the classifications of low income (18 states), middle-income (19), upper-middle income (6) and high income (7).[7] Of those 50, only 27 report macroeconomic data to the IFS tapes of the IMF and our analysis is therefore restricted to those states.[8] This means that we have a somewhat limited sample and that the strength of our conclusions will be somewhat diluted.

The basic statistics for small states are shown in the Annex Table 7.4. All the countries apart from St Kitts have had positive dollar income growth in the decade from 1985 but over a *third* have experienced a fall in dollar income *per head*. Over the period 1987–95 the average national price level has remained at a third of that of the US but there have been large within country variation for the richer states: Bahamas, Iceland, Ireland, Singapore. The national price level is – as is found for non-small states – increasing in income per head. Apart from Uruguay, these are neither high inflation countries (excluding Uruguay the average is 5.7 per cent) nor closed economies (the average ratio of exports and imports to GNP is 112 per cent). Given the wide dispersion of income per head there is a correspondingly wide dispersion of social indicators. The national accounts data are not particularly good in general, but it seems to be the case that the service sector dominates – in only CAR, Equatorial Guinea, Gambia and Togo does agriculture account for a higher percentage of GDP than services.

So how have these economies set their nominal exchange rate regimes to deal with, in aggregate: (i) highly variable per head income growth; (ii) evolving equilibrium levels for the real exchange rate; (iii) openness; and (iv) increasing importance of the tertiary sector? The model described in the section on 'Theory' would suggest by floating their exchange rate, however, Table 7.1 suggests that in small states largely choose to fix the nominal exchange rate.[9] The fact that the small states we examine have largely adopted fixed exchange rates may be related directly to the discussion in the section on theory. For small economies, the network externalities arising from the use of a currency may actually be promoted if the domestic currency is linked to a convertible currency such as the US dollar or the French franc, see Buiter (1995) on this point. Unlike the secular trend towards floating exchange rates described above, there has also been relatively little change in exchange rate regime over the period for small states. For comparative purposes

Table 7.1 Exchange rate regime in small states

Country	*1975*	*1985*	*1997*
Antigua		XCD2.70=USD1.00, peg	XCD2.70=USD1.00, peg
Bahamas	BSD1.00=USD1.00, peg	BSD1.00=USD1.00, peg	BSD1.00=USD1.00, peg
Bahrain	BHD0.395=USD1.00	BHD0.476=SDR1.00, peg	BHD0.476=SDR1.00, peg
Belize		BZD1.00=USD0.50, peg	BZD1.00=USD0.50, peg
CAR	XAF1.00=FRF0.02, peg	XAF1.00=FRF0.02, peg	XAF1.00=FRF0.01, peg
Congo	XAF1.00=FEF0.02, peg	XAF1.00=FRF0.02, peg	XAF1.00=FRF0.01, peg
Costa Rica	CRC8.57=USD1.00, peg	CRC48.00=USD1.0, peg	Managed float
Cyprus	Managed basket, flex-peg	Managed basket, flex-peg	CYP1.00=ECU1.71, peg
Equatorial Guinea	Par to ESP, peg	XAF1.00=FRF0.02, peg	XAF1=FRF0.01, peg
Fiji	FJD0.80=USD1.00	Weighted basket, peg	Weighted basket, peg
Gabon	XAF1=FRF0.02, peg	XAF1.00=FRF0.02, peg	XAF1.00=FF0.01, peg
Gambia	GMD4.00=GBP1.00, peg	GMD5.00=GBP1.00, peg	Independent float
Iceland	ISK118.00=USD1.00, peg	Weighted basket, peg	Weighted basket, peg
Ireland	Par to GBP, peg	Co-operative peg	Co-operative peg
Lesotho	(South African) Rand Monetary Area	LSL1.00=ZAR1.00, peg	LSL1.00=ZAR1.00, peg
Malta	Managed float	Weighted basket, peg	Weighted basket, peg
New Zealand	Managed float	Managed float	Independent float
Panama	PAB1.00=USD1.00, peg	PAB1.00=USD1.00, peg	PAB1.00=USD1.00, peg
Samoa (Western)	WST1.00=USD1.68, peg	Weighted basket, peg	Weighted basket, peg
Singapore	Independent float	Weighted basket, peg	Managed float
Solomon Islands		Weighted basket, peg	Weighted basket, peg
St Kitts and Nevis		XCD2.70=USD1.00, peg	XCD2.70=USD1.00, peg
St Lucia		XCD2.70=USD1.00, peg	XCD2.70=USD1.00, peg
St Vincent		XCD2.70=USD1.00, peg	XCD2.70=USD1.00, peg
Togo	XAF1=FRF0.02, peg	XAF1.00=FRF0.02, peg	XAF1.00=FRF0.01, peg
Trinidad	TTD4.80=GBP1.00	TTD2.40=USD1.00, peg	Independent float
Uruguay	Independently float	Independent float	Managed float

Note: The gaps in the first column are awaiting replies from the IMF.

we shall treat Bahrain, Costa Rica, Gambia, Ireland, New Zealand and Singapore as floating exchange rate countries.

Unconditional variability

The first three columns of Table 7.2 report the unconditional coefficient of variation for the annualised quarterly changes in the nominal exchange rates (NEER), real effective exchange rates (REER), and relative inflation rates.[10] In aggregate, there appears to be little difference between nominal and real exchange rate variability, and both seem twice as variable as relative inflation. Casually, this finding would seem to support the sticky price view of the world with real exchange rate variability closely associated with nominal exchange rate variability. This finding seems sharpened when we split the sample and find that it is particularly the floaters that have high nominal exchange rate variability. But as it is the fixed exchange rate states which have higher real exchange rate and relative inflation variability something else may be going on here.

Table 7.2 Unconditional aggregate variability and decomposition of the variability of the real effective exchange rate

Country	*Coefficient of variation*			*Variance (NEER)*	*Variance (inflation)*	*Two covariants (NEER, inflation)*
	NEER	*REER*	*inflation*			
Bahrain	52.45	3.72	0.95	0.06	0.62	0.15
Belize	1.10	24.11	4.67	0.15	0.31	0.16
CAR	5.07	4.19	16.24	0.38	3.22	–0.82
Congo	1.07	5.53	0.74	2.13	3.19	–1.22
Costa Rica	2.32	8.22	0.79	2.07	4.98	–3.45
Cyprus	1.33	2.36	3.43	0.03	0.04	0.00
Equatorial Guinea	5.47	1.43	2.11	1.66	1.66	–2.15
Fiji	13.10	5.89	1.76	0.07	0.74	–0.06
Gabon	78.63	3.73	6.67	0.92	2.81	–2.00
Gambia	7.42	6.32	1.64	1.42	3.63	–3.65
Iceland	1.19	12.65	1.00	2.84	2.99	–5.09
Lesotho	1.41	7.50	0.48	0.21	0.19	0.05
Malta	4.34	6.85	2.09	0.20	0.25	0.19
New Zealand	4.68	7.50	1.60	0.19	0.64	–0.19
Panama	0.00	1.42	0.65	0.41	0.07	–0.31
Samoa (Western)	1.50	21.67	1.14	0.46	0.36	–0.19
Solomon Islands	1.13	5.07	0.62	0.15	0.69	–0.10
St Kitts	3.06	7.08	2.40	0.04	0.15	0.00
St Lucia	2.67	16.95	27.63	0.12	0.33	–0.09
St Vincent	1.75	22.00	50.11	0.06	0.19	0.01
Togo	35.10	3.47	8.50	0.86	3.02	–2.09
Trinidad	4.92	16.78	0.62	0.24	1.62	–0.04
Uruguay	5.27	13.39	0.41	3.79	5.19	0.63
All	10.22	9.04	5.92	0.80	1.60	–0.88
Floaters	14.43	7.83	1.08	1.51	3.01	–1.30
Fixed	9.05	9.37	7.27	0.61	1.21	–0.76

Note: The nominal exchange rate series and relative inflation series have been scaled by the real exchange rate – so we measure variability per unit of real exchange rate variability.

The final three columns of Table 7.2 dig a little deeper into the previous results by decomposing the variation in the real exchange rate into parts attributable to nominal exchange rate variation, relative inflation variation and (twice) the covariation between the two. Here the results seem less supportive of the sticky price view. We find in aggregate that relative inflation variation is greater than nominal exchange rate variation; and only in the cases of Samoa, Panama, and Lesotho is nominal exchange rate variability greater. We also find that relative inflation rates and nominal exchange rates negatively covary in 15 out of 23 cases; suggesting that there is, in the majority of these countries, some immediate change in prices acting to cushion the impact of nominal exchange rates on real exchange rates.[11]

And in comparing floaters to fixers we find that any increase in the variance of real exchange rates in floating regimes is linked to greater variability of *both* nominal effective exchange rates and relative inflation rates – it therefore does not appear to *result* from the choice of nominal exchange rate.

Spectral decomposition

Finally, Table 7.3 takes this analysis one stage further and decomposes the real effective exchange rate and its two constituents, using spectral analysis, into orthogonal series each identified with different frequencies. Concentrating on the estimated spectra for the real effective exchange rate, we find no significant differences in the proportions of the overall unconditional variation explained by different frequencies for either floaters or fixers. Recall that a sticky-price model would lead to high variability in high frequency spectra because a nominal shock would lead to an equivalent shock to the real exchange rate that would dissipate through time. Moreover the pattern of real exchange rate innovations which emerge is not consistent with a fixed real exchange rate equilibrium – the lowest frequency component explains nearly a fifth of innovations. Note that stationarity has been induced by the first difference operator so the weight of high frequency components is designed to be high in this experiment. Even then it is the lowest frequency for the real exchange rate which tends to be the most important: suggesting the existence of important equilibrium shifts even over a period as short as 20 years. Finally, neither the nominal exchange rate nor relative inflation rates alter their spectral pattern with respect to each other or the choice of regime. We cannot then even ascribe the choice of floating as an endogenous to states with stickier prices. Flexible prices just seem to be part of the story for these small states.[12]

We put most weight on the results of the identified spectra. These results suggest quite clearly that the sticky price view of the world – as far as small states are concerned looks problematic. What behavioural stories could explain the lack of price stickiness? We suggest four: (i) a very high proportion of traded goods mean that the exchange rate forms a large weight in the domestic price index and any form of mark-up pricing is likely to reference the exchange rate; (ii) the importance of the service sector may help explain price and wage flexibility; (iii) contracts or institutional practices may tend towards higher frequency negotiations of wages; and (iv) connected with the two previous points, the possibility of rigidities caused by monopolistic competition may be reduced in such an economy.

The implications of these results is that real exchange rate movements seem to be equilibrium phenomena – not Dornbusch-style overshooting phenomena – but because prices are flexible the choice of exchange rate regime does not hinge on the question of economic stabilisation. Equally the choice of exchange rate regime does not *per se* impact on the variability of the real exchange rate. The lack of any *a priori* welfare loss from a fixed exchange rate means that the preference for maintaining exchange rate fixity in some form – peg or target zone – reflects a policy optimisation choice resulting from somewhere outside the credibility-flexibility paradigm. This is because weakness in monetary policy credibility cannot arise when there are no incentives to generate policy shocks.

Table 7.3 Variance proportion by cycle length

Country		*Lowest*	*4 years≤*	*2 years≤*	*1 year≤*	*2 quarters≤*	*1 quarter≤*
Bahrain	*n*	27.42	25.39	17.66	19.49	6.49	3.53
	r	33.10	28.52	16.94	15.33	5.32	0.79
	p	33.21	28.65	15.08	13.39	4.50	5.17
Belize	*n*	18.02	22.76	21.62	19.04	9.15	9.42
	r	26.07	22.77	16.79	21.51	8.78	4.08
	p	9.20	13.87	17.76	25.17	15.26	18.73
CAR	*n*	13.46	16.29	18.79	25.34	9.28	16.83
	r	11.49	13.72	17.42	26.91	10.50	19.96
	p	27.86	25.40	19.28	20.47	4.80	2.19
Congo	*n*	38.52	36.38	18.45	2.87	1.22	2.56
	r	6.60	11.37	18.32	18.56	11.95	33.21
	p	39.47	29.78	15.53	8.34	3.39	3.49
Costa Rica	*n*	21.06	23.97	23.21	19.66	2.26	9.84
	r	9.22	14.35	20.93	31.86	4.62	19.02
	p	20.08	25.90	26.87	19.80	1.94	5.41
Cyprus	*n*	12.87	19.35	23.75	29.20	11.50	3.34
	r	32.47	24.00	13.94	18.09	7.92	3.58
	p	26.10	22.05	14.52	16.05	9.95	11.32
Equatorial Guinea	*n*	13.44	15.11	17.54	25.20	15.22	13.49
	r	16.61	17.45	17.51	22.18	11.38	14.86
	p	22.83	23.43	16.85	12.76	9.33	14.79
Fiji	*n*	22.19	23.74	19.80	21.91	10.21	2.16
	r	29.34	26.45	16.25	17.01	8.94	2.02
	p	22.03	22.06	19.65	23.34	7.69	5.22
Gabon	*n*	20.37	19.37	17.84	24.98	9.13	8.30
	r	9.16	13.02	21.73	33.41	12.08	10.58
	p	20.40	23.09	22.45	24.06	7.63	2.37
Gambia	*n*	20.25	20.05	17.33	26.96	12.28	3.13
	r	4.52	8.67	15.29	34.88	24.47	12.17
	p	35.34	28.11	14.53	14.55	5.73	1.74
Iceland	*n*	42.91	28.52	12.34	12.40	2.63	1.21
	r	9.85	16.09	23.15	31.25	11.71	7.94
	p	51.00	30.99	8.09	6.26	2.54	1.11
Lesotho	*n*	38.11	24.12	9.72	14.55	9.82	3.67
	r	18.51	18.91	17.99	23.09	15.40	6.11
	p	24.86	23.48	19.76	17.65	9.58	4.67
Malta	*n*	35.80	27.53	15.67	13.79	5.51	1.70
	r	20.59	17.48	11.23	14.74	13.86	22.09
	p	37.17	30.22	14.47	10.36	4.48	3.30
New Zealand	*n*	23.68	18.74	11.33	25.18	13.63	7.43
	r	19.12	20.54	15.67	23.07	14.20	7.41
	p	40.06	29.70	14.18	11.28	3.16	1.62
Panama	*n*	0.00	0.00	0.00	0.00	0.00	0.00
	r	38.33	28.86	13.82	10.99	4.16	3.83
	p	29.08	24.12	15.57	18.81	5.94	6.47
Samoa (Western)	*n*	40.26	25.55	8.87	8.99	11.86	4.48
	r	20.62	16.48	11.17	29.03	16.74	5.95
	p	18.53	14.77	11.23	34.66	16.45	4.37

(Continued …)

Table 7.3 (continued)

Country		*Lowest*	*4 years≤*	*2 years≤*	*1 year≤*	*2 quarters≤*	*1 quarter≤*
Solomon Islands	*n*	32.84	27.78	18.41	13.96	4.72	2.28
	r	24.31	22.60	19.19	22.45	8.77	2.68
	p	24.29	18.46	11.72	25.41	12.46	7.66
St Kitts	*n*	17.95	17.81	16.04	29.75	12.79	5.67
	r	21.42	18.33	13.37	27.16	13.43	6.29
	p	24.51	22.84	16.55	21.34	10.22	4.54
St Lucia	*n*	21.90	22.97	19.03	23.57	8.52	4.00
	r	20.90	16.94	13.81	31.84	10.17	6.33
	p	20.98	24.43	23.67	20.79	4.46	5.67
St Vincent	*n*	14.94	18.01	19.83	29.75	11.66	5.82
	r	18.20	17.29	16.13	29.39	11.94	7.05
	p	22.47	20.32	17.91	21.07	9.27	8.96
Togo	*n*	18.53	18.62	18.41	26.70	9.76	7.97
	r	6.88	11.39	19.30	36.48	14.53	11.43
	p	27.64	25.79	19.31	19.03	5.24	2.98
Trinidad	*n*	26.56	25.03	17.81	18.37	7.89	4.35
	r	26.85	25.63	18.03	17.76	6.97	4.75
	p	15.14	22.50	26.88	21.47	10.36	3.66
Uruguay	*n*	18.17	23.75	26.11	22.97	6.18	2.82
	r	22.36	22.76	18.04	25.30	7.59	3.95
	p	42.94	32.42	13.53	5.28	2.80	3.03
All	*n*	24.51	22.77	17.71	20.66	8.71	5.64
	r	19.41	18.85	16.78	24.45	11.11	9.40
	p	27.62	24.45	17.19	17.88	7.27	5.59
Floaters	*n*	22.12	22.38	19.13	22.85	8.17	5.35
	r	17.66	18.97	17.37	26.09	11.24	8.67
	p	34.33	28.96	16.84	12.86	3.63	3.39
Fixers	*n*	25.22	22.88	17.29	20.02	8.87	5.72
	r	19.90	18.82	16.62	23.99	11.07	9.60
	p	25.75	23.20	17.29	19.28	8.28	6.19

Notes: The variance proportions are given by the average of estimated spectral density functions for approximations using each of Bartlett, Tukey and Parzen windows. Estimates were undertaken using the annualised first difference in quarterly data, where *n* is changes in the nominal effective exchange rate, *r* is changes in the real effective exchange rate and *p* changes in relative inflation rates. Standard errors of the estimates are available on request.

Conclusion

Small states tend to fix, relatively rarely do they float.[13] There is now mounting evidence that real asymmetric shocks seem to be a significant feature of open economies, see Chadha and Hudson (1998) for example. This chapter further suggests that there are important longer run components in real exchange rate developments, which seem incompatible with fix price stories. It appears that small states seem prone to considerable output variability and yet most of these economies maintain fixed exchange rates. This is because as prices in these states do not seem particularly sticky, there is little gain from using domestic monetary policy to stabilise output. For Romer (1983) type reasons therefore there is also then

little inflation bias produced by an independent monetary authority because the gains from unexpected inflation shocks under these circumstances are very limited. Equally under these circumstances real exchange rate adjustment can occur just as easily by changes in the domestic price level and so exchange rate fixity can freely be maintained.

Such a stylised fact should make the monetary authorities indifferent between fixed and floating exchange rates in terms of output stabilisation. So the critical issues facing a small state in terms of the choice of exchange rate regime, abstracting from some degree of inertia, are likely to involve political economy concerns such as sharing in the provision of public goods. For example, formal external exchange rate links may encourage financial regulation, which may be important in offshore financial centres such as Malta, or help to maintain political ties with some former colony or large neighbour, for example in the CFA zone or the Western hemisphere states. What comes to mind is that the small states do not play the credibility–flexibility game with their currencies because they have no need to do so.

But because such countries are unlikely to have the capabilities to develop a number of the pieces of the financial jigsaw, for example, an independent central bank, financial regulation and deep money markets, they have to draw some formal links with a larger nation. What goes for micro-states may to some extent also then apply to other developing countries in general. That is the choice of exchange rate regime cannot be nested within the optimal stabilisation debate that is rightly used to frame the issue in advanced countries. If part of the picture of lesser development is also a lack of political and institutional capability then the association with best practice of the developed world, through something as symbolic as the exchange rate, may provide an important benefit. Developing a theoretical model, which can deal with these questions with respect to issues of currency union, will be extremely welcome.

Acknowledgements

I am grateful to participants of the International Economics Study Group Conference on ‘Small states in the International Economy’ at Birmingham University in April 1998 for comments and to Prasanna Gai, Stephen Millard, Gabriel Sterne and especially Mansoob Murshed and Robert Read for a constructive interchange of thoughts.

Notes

1 Of course, Marston (1985) pointed out that, strictly speaking, it is the relative variance of real (IS) shocks to nominal (LM) shocks which increases the case for floating exchange rates.
2 Recall both developing and advanced economies were members of the gold standard and the Bretton Woods system.
3 Source: IMF *World Economic Outlook* October 1997.

Annex Table 7.4 Basic statistics of small states

Country	*Population level 1995*	*Per year growth 1985–95*	*Income/ head($) 1995*	*Income/ head growth 1985–95*	*PPP % US 1987*	*1995*	*Nominal er % of PPP 1995*	*Infant mortality per 1000 live*	*Average inflation 1985–95*	*Trade (openness)*	*Agriculture % of GDP*	*Services % of GDP*
Antigua	65	0.5	3–9000	2.7				18	4.4	217	4	22
Bahamas	276	1.7	11940	–1.0	68.8	54.5	81.17	15	3.2			
Bahrain	577	3.1	7840	0.6	54.2	49.7	58.51	19	0.4	191	1	27
Belize	216	2.6	2630	4.4	17.0	20.0	48.70	36	3.5	109	20	26
CAR	3275	2.3	340	–2.4	5.0	4.0	31.78	98	3.7	46	44	15
Congo	2633	3.1	680	–3.2	11.5	7.6	33.17	90	2.2	128	10	27
Costa Rica	3399	2.5	2610	2.8	19.8	21.7	44.62	13	18.5	81	17	25
Cyprus	734	1.2	*c.* 10000	4.6	44.5			8	4.3	99	5	22
Equatorial Guinea	400	2.5	380	2.3				111	4.1	112	50	23
Fiji	775	1.1	2440	2.3	19.1	21.4	42.21		4.9	104		14
Gabon	1077	2.9	3490	–8.2	n/a	n/a		89	4.8	101		26
Gambia	1113	4.0	320	0.3	4.5	3.5	34.41	126	10.3	103	28	21
Iceland	268	1.1	24950	0.3	88.1	75.8	121.95	4	11.8	70		15
Ireland	3586	0.1	14710	5.2	44.2	58.1	93.81	6	2.5	136		13
Lesotho	1980	2.4	770	1.2	6.1	6.6	43.26	76	13.6	138	10	87
Malta	372	0.8	3–9000	5.1	38.2		45.45	9	2.9	198	3	29
NZ	3601	1.0	14340	0.8	63.3	60.6	87.65	7	3.9	62		24
Panama	2631	1.9	2750	–0.4	26.1	22.2	45.99	23	1.7	79	11	24
Samoa (Western)	165	0.5	1120	–0.4	8.9	7.5	55.17	22	10.6			
Singapore	2987	1.8	26730	6.2	56.1	84.4	117.39	4	3.9		0	33
Solomon	375	3.1	910	2.2	7.9	8.1	41.55	41	11.7			
St Kitts	41	–0.4	5170	4.6	28.2	34.9	54.94	31	5.5		6	39
St Lucia	158	1.4	3370	3.9				17	3.2	141	11	25
St Vincent	111	0.8	2280	3.9				19	3.6		11	
Togo	4085	3.0	310	–2.7	5.5	4.2	27.43	88	3	65	38	14
Trinidad	1287	0.9	3770	–1.7	38.1	31.9	43.79	13	6.8	68	3	14
Uruguay	3184	0.6	5170	3.1	23.6	24.6	77.98	18	70.5	41	9	14

4 There is no advantage to tying ones hands in the sense of Giavazzi and Pagano (1988) if those hands cannot be put to any misuse anyway.
5 This argument echoes Romer's (1993) finding of a negative relationship between openness and inflation because of a lower inflation bias in such economies: the lower inflation bias makes the fixed exchange rate less attractive to policy-makers.
6 We might want to think of the relative weight on output losses to those of inflation as being exogenous. But an independent monetary authority might be charged with pursuing policy in such a form as to ensure indifference between a fixed and floating exchange rate.
7 Under high income, I have included Ireland and New Zealand which have 1995 populations of 3.6 million, but as it is not possible to be sure of the steady-state populations from one year's reading they are retained in the sample.
8 In fact, of the 27, Antigua (no domestic CPI series), Bahamas (incorrect real effective exchange rate series), Ireland and Singapore (absent real effective exchange rate series) are omitted in much of the analysis.
9 The source is IMF's *Annual Reports on Exchange Rate Arrangements*.
10 The data is sourced from the IMF's IFS tapes and covers the period from 1979(1) to 1997(3).
11 Some of this cushioning may result from a direct price level effect on tradables arising from the high degree of structural openness.
12 Further evidence, for example in Armstrong and Read (1998) suggests that this observation is reasonably robust.
13 If they were truly indifferent between the two, we might expect the states to be split 50:50.

References

Alesina, A. and E. Spolaore (1997) 'On the number and size of nations', *Quarterly Journal of Economics*, 112: 1027–56.

Alesina, A., R. Perotti and E. Spolaore (1995) 'Together or separately? Issues on the costs and benefits of political and fiscal unions', *European Economic Review*, 39: 751–8.

Armstrong, H.W. and R. Read (1998) 'Trade and growth in small states: the impact of global trade liberalisation', *World Economy*, 21 (4): 563–85.

Buiter, W.H., (1995) *Macroeconomic Policy During a Transition to Monetary Union* CEPR Working Paper no. 1222, London.

Chadha, J.S. and S.L. Hudson (1998) *The OCA Case for EMU: A Structural VAR Approach*, Discussion Paper in Economics and Econometrics, Department of Economics, Southampton University.

Devarajan, S. and D. Rodrik (1991) *Do the Benefits of Fixed Exchange Rates Outweigh the Costs? The Franc Zone in Africa* CEPR Working Paper no. 561, London.

Giavazzi, F. and Pagano, M. (1988) 'The advantage of tying one's hands: EMS discipline and central bank credibility', *European Economic Review*, 32: 1055–82.

Goodhart, C.A.E. (1995) 'The political economy of monetary union', in P. Kenen (ed.) *Understanding Interdependence*, Princeton PA: Princeton University Press.

IMF, *Exchange Arrangements and Exchange Restrictions Annual Report*, Various, International Monetary Fund: Washington DC.

Kenen, P. (1969) 'The theory of optimum currency areas: an eclectic view', in R.A. Mundell and A. K. Swoboda (eds) *Monetary Problems of the International Economy*, Chicago: University of Chicago Press.

Marston, R. (1985) 'Stabilisation policies in open economies,' in R.W. Jones and P.B. Kenen (eds) *Handbook of International Economics*, Vol. 2, Amsterdam: North-Holland.

Mundell, R.A. (1961) 'A theory of optimum currency areas', *American Economic Review*, 51: 657–664.

Romer, D. (1993) 'Openness and inflation: theory and evidence', *Quarterly Journal of Economics*, 108: 869–903.
World Bank, *World Tables*, Various, Washington DC: World Bank.

8 Trade in fragmented products

Concepts and evidence

Holger Görg

'Fragmentation' of production refers to the splitting up of previously integrated production processes into various components, allowing trade in formerly non-tradeable components and offering new possibilities for international specialisation of production (Jones and Kierzkowski, 2001). Fragmentation, thus, not only increases scope for international trade, but also for the emergence of multinational companies (MNCs), with plants in different countries producing different components which are then traded between plants within the multinationals.

While fragmentation has attracted considerable interest in the recent literature (see, for example, Arndt, 1997; Deardorff, 1998, 2001; Jones and Kierzkowski, 2001), much of the literature to-date has discussed theoretical issues from the point of view of international trade theory using Ricardian and Heckscher–Ohlin–Samuelson (HOS) type models. For example, Jones and Kierzkowski (1990, 2001) discuss the importance of service links for the production of separate fragmented components. The production of separate components needs to be co-ordinated, using transportation and communication services, and reductions in costs for these services have facilitated fragmentation. Harris (1995, 2001) concentrates further on the importance of telecommunications for fragmentation, while Deardorff (1998, 2001) analyses the determination of world prices for fragmented components.

Thus, these papers implicitly relate to important political economy issues related to fragmentation. For example, the importance of services links for the breaking up of production processes implies that further liberalisation of services trade under the General Agreement on Trade in Services (GATS) within the new WTO system may have positive effects on the growth of fragmentation. Also the resulting new patterns of specialisation around the globe and the determination of prices for fragmented products have important policy implications, not least for developing countries.

Complementing trade theory, this chapter briefly presents an alternative way of looking at fragmentation, namely, transaction cost economics. Fragmentation poses a question of 'make or buy', whether a firm should produce components itself or source them from outside. We also undertake an empirical study of the extent of fragmentation in US trade with the EU, to contribute some empirical

work to the, so far, mainly theoretical debate on fragmentation. Data available from Eurostat enable us to analyse US trade with the EU in intermediate goods, which are processed in the EU and then re-exported to destinations outside the EU. This type of trade, which we use as a proxy for the extent of fragmentation, is referred to as 'inward processing trade'.

This chapter is structured as follows. The first section briefly surveys the previous work on fragmentation from the point of view of international trade theory and presents a transaction costs approach to looking at fragmentation. The second section discusses how one may measure fragmentation and presents some descriptive data pertaining to the extent of US inward processing trade in the EU. The third section presents an econometric estimation of the determinants of US inward processing trade, while the final section summarises the main results and points out some directions for further research.

Theoretical concepts

International trade theory

Most of the literature concerned with fragmentation thus far has attempted to explain this phenomenon from the point of view of international trade theory based on Ricardian and HOS type trade models. Figure 8.1, taken from Jones and Kierzkowski (2001) illustrates the issues involved in fragmentation. Assume that a country which engages in international trade produces good *I* in an initial equilibrium using two factors, capital and labour in fixed proportions. Point *I* indicates the input combination necessary to produce one dollar's worth of this product at going world prices. Product *I* is composed of two intermediate goods (fragments) *A* and *B* though, initially, technology does not allow to split the production process into *A* and *B*.[1] There is, therefore, no market price for components *A* and *B*, though the costs of producing the components can be imputed giving knowledge of the production technique (Jones and Kierzkowski, 2001). Points *A* and *B* represent the input combinations necessary to produce one dollar's worth of each component, based on their implicit cost of production, and point *I* represents a weighted average of these input requirements.

Fragmentation allows the breaking up of the production process, i.e., components *A* and *B* can be produced separately. The fragments can be traded directly on world markets and explicit world prices will be determined.[2] This opens up new possibilities for specialisation, as there are a larger number of tradable goods than before fragmentation. The production of components can be expected to disperse globally according to countries' comparative advantage.

For the hypothetical country in Figure 8.1, comparative advantage would point towards specialisation in the production of component *A* since it takes less of both factors to produce one dollar's worth of *A* than of *B*. This, of course, is only true if the relative prices of *A* and *B* remain the same as before fragmentation, which one would not expect. As Jones and Kierzkowski (2001) point out, one may assume that prices for each segment falls after fragmentation as a result of international

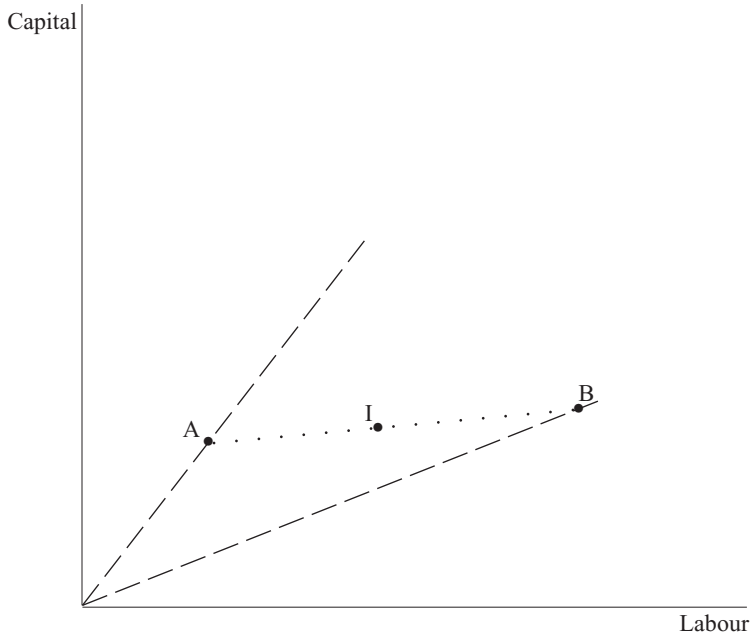

Figure 8.1 Fragmentation in a simple trade model

competition and specialisation. If the relative prices of *A* and *B* change it may become more profitable to produce *B* than *A* for the country. Most importantly, however, fragmentation can be expected to lead to a reduction in costs of obtaining the combination of fragments *A* and *B* which are needed to produce the integrated good *I* due to specialisation.

There are a number of possible reasons to explain fragmentation. Harris (2001) argues that fragmentation is driven by the search for gains from specialisation. Specialisation is limited by the existence of global co-ordination costs, such as costs for communication and transport. Hence, Harris suggests that the improvements in communication, such as fax and internet, as well as reductions in transport costs have considerably extended the scope for global specialisation and fragmentation (see also Harris, 1995 and Jones and Kierzkowski, 2001). Arguably, trade barriers also pose a limit on international specialisation, and the reduction in trade barriers following GATT and WTO should also increase the potential for fragmentation.

There has been some discussion of the effects on the demand for skilled and unskilled labour in the countries involved of this specialisation due to fragmentation. If fragmentation leads to the relocation of labour-intensive low-skill production into labour-abundant countries (which are mainly developing countries) and capital-intensive high-skill production into capital-abundant countries (mainly developed countries), there will be a shift in labour demand away from low-skilled towards high-skilled labour in developed countries. This shift in labour demand may lead to increases in wage inequality as wages for low-skilled labour will

decrease relative to high-skilled labour, or it may lead to increases in unemployment amongst low-skilled labour.[3]

Transaction cost economics

Of course, fragmentation in essence is about the question whether a firm should make-or-buy intermediate inputs; a question that has a long tradition in industrial organisation and, more specifically, transaction cost economics. Another way of looking at fragmentation, therefore, is from the point of view of transaction cost economics (Coase, 1937, 1992; Williamson, 1975, 1979). This is not to say that this is a substitute for the international trade theory approach, rather it should be seen as complementing that literature. The international trade approach stresses the effects of international fragmentation on, for example, changes in comparative advantage and relative prices while transaction cost economics focuses on processes within individual firms.

From the point of view of transaction cost economics, the possibility of splitting up the production process leads to the question as to whether a plant should produce components within its own plant, or source them from outside, see Lyons (1995) and Liu and Yang (2000). Fragmentation will be profitable if the costs of sourcing intermediate inputs from outside the plant (including production and transaction costs) are lower than the costs of producing the inputs within the plant. Thus, transaction cost economics emphasises the firm-level perspective of fragmentation.

To illustrate the transaction costs approach, consider the following very simple case. Good Y is produced using a linear combination of the components A and B and, for simplicity, assume that the components are produced only with a single homogenous factor of production, labour. Total costs for producing Y are then $C_Y = c(W_A, W_B)$ where W is the total factor cost of component A and B respectively. The components can be produced within the plant or can be bought from outside the plant. Marginal costs of producing an additional unit of the component are $w_x = w_{ix}$ (where $x = A, B$) if the component is produced within the plant, or $w_x = w_{ox} + s$ if it is produced outside and shipped to the plant for production of the final product Y, where s denotes additional transaction costs due to the sourcing of the product outside the plant. It will be profitable for a plant to produce the components within the plant if $w_{ix} < w_{ox} + s$.

This condition simply re-states the equilibrium condition for vertical integration by an individual plant given by Coase (1937) who argued that

> a firm will tend to expand until the costs of organizing an extra transaction within the firm becomes equal to the costs of carrying out the same transaction by means of an exchange on the open market or the costs of organizing in another firm (p. 395).

If s is high it may not be profitable for a plant to source inputs from outside and it may prefer to produce the components within its own plant at marginal costs w_{ix}. The contrary is the case if s is very low and $w_{ox} < w_{ix}$.

To examine the issues involved more closely it is useful to introduce two conceptual distinctions at this stage. First, fragmentation can be intra-firm or inter-firm. In the case of the former, plants belonging to the same firm specialise in the production of different components. In the case of the latter, different firms engage in such specialisation. Second, fragmentation can take place across borders or within the same country. Intra-firm fragmentation across borders leads to the emergence of multinational companies, while inter-firm fragmentation across borders creates inter-firm trade.

While international fragmentation arguably offers more possibilities for exploiting differences in factor costs across countries (w in the model above), fragmentation within the same economy may, on the other hand, involve lower transaction costs. These costs include the direct transaction costs of fragmented production, such as costs for transport and communication, as well as the costs of potential undesired spillovers of proprietary capital. Furthermore, differences in culture, language, and legal systems may impose additional costs on the firm engaging in fragmentation (Jones and Kierzkowski, 2001).

As regards intra- versus inter-firm fragmentation, a firm will choose between either alternative in response to the particular transaction costs involved. In the case of intra-firm fragmentation, the transaction is internal to the firm that precludes any negative external market effects which could occur due to, for example, the transfer of technology to, and subsequent exploitation by, the external firm producing the component. However, intra-firm fragmentation may involve diseconomies of scope; the attempt to produce all necessary components within the same firm may lead to higher production costs. Also, sunk costs involved with setting up an own plant in another location may be substantial, such that lower factor costs in different locations may not be fully exploited through intra-firm fragmentation. In the case of inter-firm fragmentation, a firm is able to source components from different locations (national and international) and may, thus, be able to reap the benefits of specialisation more fully. However, the firm will have to incur the additional transaction costs involved in inter-firm fragmentation.

As pointed out above, the recent technical progress in the services sector, which has led to the tradeability of many services and to reductions in the costs of international communications suggest that we may assume transaction costs s to be falling, thus facilitating fragmentation (see Harris, 2001; Jones and Kierzkowski, 2001). Also, the reduction of trade barriers for manufacturing and services trade within the WTO can be expected to reduce s and allow greater scope for fragmentation. If s is negligible, fragmentation will always be profitable when $w_{ox} < w_{ix}$, which can be assumed to be the case if the production outside the plant takes place in, for example, a lower wage-cost location.

Regardless of the type of fragmentation chosen by the firm, this brief discussion suggests that, if transaction costs are low, production of components is likely to move to the location with lowest factor costs. In a general equilibrium framework with more than one factor of production this simple finding generalises to the result that the production of the components move to the country with a comparative advantage in the production of those components.

Measuring fragmentation

For the measurement of fragmentation, we limit the analysis to international trade as a means of fragmentation. Numerous studies, both for single countries and for groups of countries, have analysed the increasing significance of foreign direct investment and multinational companies (see the survey in Caves, 1996) as well as intra-firm trade within multinationals (see, for example, Helleiner, 1981 and Svensson, 1996). Empirical studies of inter-firm trade as a means of fragmentation, however, seem to have been less prevalent in the literature. There have been only a few empirical studies of the importance of trade as a means of fragmentation, such as Feenstra (1998), Baldone et al. (2001), Kierzkowski (2001), Ruane and Görg (2001) and Yeats (2001).

One indicator of trade in fragmented components (rather than final goods) is the extent of trade which is devoted to trading intermediate goods which are processed abroad and are then shipped back to the home or a third country for final production (Kierzkowski, 2001). Such data are available for the EU, where Eurostat collects information on outward and inward processing trade (OPT/IPT) in EU member states.[4] Outward processing is the customs arrangement allowing goods to be temporarily exported from EU territory for processing, and the resultant products to be released for free circulation in the EU with total or partial relief from import duties.[5] Inward processing is the duty relief procedure allowing goods to be imported into the EU for processing and subsequent export outside the EU without payment of duty. Thus, IPT data enable us to establish to what extent countries in the EU are bases for fragmentation and they, therefore, give a good indication of the extent of fragmentation across borders. Unfortunately, the data do not allow us to distinguish between intra-firm trade, i.e., trade between different plants within a multinational company, and inter-firm trade, i.e., trade between different firms.

We focus on the US as a source for IPT because the US is a major trade partner of EU countries. According to Eurostat trade data, trade with the US accounted for roughly 18 per cent of total external trade of the EU in 1994.[6] Also, the US is a major investor in the EU, accounting for roughly one-third of total foreign direct investment stocks in the EU in the early to mid-1990s (Dunning, 1997).[7] Furthermore, as with many empirical studies, constraints on the data used for the empirical estimation below favour the focus on the US as a source for fragmentation.

Table 8.1 shows the importance of IPT in terms of total imports from the US. Roughly 20 per cent of imports from the US into the EU were IPT, i.e., roughly one-fifth of total imports from the US were in fragmented components. The importance of US IPT is in line with the findings of empirical studies of intra-industry trade (IIT), which find that a large share of international trade is trade within industries and that IIT is mostly vertical rather than horizontal (Greenaway and Milner, 1986; Greenaway and Torstensson, 2000). Trade in fragmented components is, arguably, one particular form of vertical intra-industry trade.

The share of IPT received varies across member states, with IPT being most important for Ireland and France. The data show that Greece and Portugal receive only very low levels of trade in fragmented components. This seems to suggest

Table 8.1 US IPT as a percentage of US imports by country, 1994

Belgium/Luxembourg	16.4
Denmark	11.3
France	30.1
Germany	16.3
Greece	3.9
Italy	10.3
Ireland	44.1
Netherlands	13.9
Portugal	6.1
Spain	11.5
UK	18.6
EU 12	19.4

Source: Calculated from Eurostat: Intra- and extra-EU trade (combined nomenclature). CD-ROM.

that these two countries are not a very attractive base for fragmented production. According to the theory outlined above, this may be due to the fact that these countries have a comparative disadvantage in production in the sectors in which US IPT in the EU takes place, or it may be due to other country or sector-specific effects. We return to this issue below.

Table 8.2 indicates that there are considerable differences in the sectoral distribution of IPT across core and peripheral countries.[8] Core countries attract IPT disproportionately in the industrial machinery and transport equipment sectors. These are sectors where increasing returns to scale are important. Peripheral countries, on the other hand, have very high rates of US IPT in the leather and non-metallic minerals sectors. We also include the data for Ireland and France in the table, which are peripheral and core countries respectively, and where IPT is disproportionately important. The differences in the sectoral distribution of IPT may again be due to different comparative advantages across member states, or they may reflect differences in production costs (particularly labour costs) in different sectors across different EU member states.

Neven (1990) shows that Ireland has a revealed comparative advantage in the production of high human capital intensive goods such as chemicals and electronics, and Table 8.2 indicates that these sectors also receive high shares of US inward processing trade. Also, the IPT data show that France receives high shares of US IPT in industrial machinery, metals, and metal manufacturing, which is in line with Neven's finding that France seems to have, to some extent, a comparative advantage in industries intensive in human capital (Neven, 1990: 26). We discuss the importance of comparative advantage in more detail below.

Table 8.2 US IPT as a percentage of US imports by sector, 1994

	EU 12	*Core*	*Periphery*	*Ireland*	*France*
Food, drink and tobacco	4.4	5.2	0.8	1.1	3.9
Chemicals	15.5	14.2	27.2	60.3	20.4
Leather	21.7	11.4	62.3	6.7	1.5
Rubber	12.4	12.3	13.0	32.9	16.9
Cork and wood	2.4	2.6	0.4	0.6	3.4
Paper	8.1	8.5	3.9	21.1	1.4
Textiles	11.0	9.6	24.9	32.3	6.1
Non-metallic minerals	6.7	4.2	53.0	67.0	2.5
Metals	19.2	19.5	16.4	17.7	29.2
Metal manufacturing	17.4	15.8	32.9	40.1	23.7
Industrial machinery	34.6	36.4	11.7	12.0	57.0
Electronics	21.0	17.5	46.2	58.1	18.4
Transport equipment	24.8	27.6	4.6	1.3	39.3
Others	10.0	9.4	16.7	33.9	9.5
Total	19.4	18.9	23.7	44.1	30.1

Source: Calculated from Eurostat: Intra- and extra-EU trade (combined nomenclature). CD-ROM.

Determinants of fragmentation[9]

Methodology

In this section we examine in some more detail the determinants of fragmentation: the factors impacting on the sectoral spread of IPT across EU countries. If the EU were one homogenous producer market, we would expect IPT to be randomly distributed across member states. If US producers, however perceive differences across locations in the EU, we would expect systematic differences in the distribution of IPT across EU countries.

As outlined in the theory above, we would, in particular, expect a member state's comparative advantage and/or the absolute level of factor costs to have an impact on its receiving IPT from the US. We would expect a positive effect of a country's comparative advantage on the level of IPT into the country. By the same token, the lower factor costs, the more IPT inflows into the country. As labour may be a particularly important factor for the production of intermediate components, we focus on labour costs in the analysis below.

As pointed out above, the IPT data may also include intra-firm trade within multinationals. It is, therefore, necessary to control for the presence of US foreign direct investment (FDI) in the country. Other things being equal, we expect the level of US FDI to have a positive impact on IPT through the effect on intra-firm trade. This leads to the following empirical model:

$$IPT_{ijt} = \alpha_0 + \alpha_1 PERFDI_{ijt} + \alpha_2 WAGE_{ijt} + \alpha_3 CA_{ijt} + \mu_i + \nu_j + \varepsilon_{ijt} \qquad (8.1)$$

where the dependent variable IPT_{ijt} is defined as the ratio of US IPT to total imports from the US in country i and sector j at time t. $PERFDI_{ijt}$ is the stock of

US foreign direct investment in country i and sector j at time t as a percentage of GDP, $WAGE_{ijt}$ is a sectoral labour cost variable and CA_{ijt} is a proxy for a country's comparative advantage in a sector. The term μ_i is a country-specific time-invariant and unobservable effect, such as culture or language, ν_j is a sector-specific time-invariant unobservable effect (for example, the level of technology), and ε_{ijt} is the remaining period-specific error term, assumed to be independent across countries, sectors, and over time.[10]

Data

Data on US IPT and comparative advantage are taken from the Eurostat CD ROM: *Intra- and Extra-EU Trade (Combined Nomenclature)* which provides intra- and extra-EU trade statistics for all member states. We calculate two measures of revealed comparative advantage, namely the net trade ratio, calculated as $CA1_{i,t} = (X_{i,t} - M_{i,t})/(X_{i,t} + M_{i,t})$ which has been used as a measure of revealed comparative advantage in recent papers by Lundbäck and Torstensson (1998) and Milner and Pentecost (1996). Also, we calculate the simple ratio of exports over imports as used by Driffield and Munday (2000) as a measure of revealed comparative advantage.[11] We calculate these two indices also using the Eurostat trade statistics.

Data on labour costs are available by sector from data from the US Department of Labor. They report hourly compensation costs for production workers in manufacturing for 31 countries and 40 manufacturing industries. We use an index for relative labour costs, i.e., labour costs in the EU country relative to the US, as a proxy for *WAGE*.[12]

Finally, the data on US foreign direct investment are taken from data available at the US Department of Commerce. The data relate to US FDI stocks in man ufacturing and services industries, which are collected by the US Department of Commerce for more than 50 countries. *PERFDI* is calculated as the ratio of US FDI stocks by sector over GDP in country i at time t, a measure also calculated in UNCTAD (1997) for most developed and developing countries. GDP data for EU member states are taken from Eurostat statistics.[13]

Econometric results

We estimate equation (8.1) with data for 14 manufacturing sectors (aggregated based on SITC Rev. 3 classification) for the period 1988 to 1994, using fixed-effects panel data regression techniques as described by Baltagi (1995). The fixed-effects model was deemed preferable for the estimation of country-specific effects to a random-effects specification because we follow a specific set of countries over time. While the fixed-effects model purges country-specific effects, we include sectoral dummies in the estimation to control for sector-specific effects as specified in equation (8.1).[14]

Columns 1 and 2 in Table 8.3 present the results of the estimation using data for all EU countries. We estimate two alternative specifications: column 1 uses *CA*1 as the measure of comparative advantage while column 2 presents results using *CA*2.

Table 8.3 Econometric results for estimation of equation (8.1)

	All EU		*EU Core*		*EU Periphery*	
Variable	*1*	*2*	*3*	*4*	*5*	*6*
CA1	0.1365	–	0.0791	–	0.0888	–
	(0.0185)†		(0.0211)†		(0.0378)‡	
CA2	–	0.0135	–	0.0329	–	0.0144
		(0.0041)†		(0.0090)†		(0.0053)†
PERFDI	7.2519	7.9046	–0.5436	–0.8448	6.7594	7.0219
	(1.2506)†	(1.2670)†	(2.6312)	(2.6439)	(1.6380)†	(1.6332)†
WAGE	0.0006	0.0007	0.0005	0.0005	0.0024	0.0025
	(0.0004)*	(0.0004)*	(0.0004)	(0.0004)	(0.0013)*	(0.0013)*
CONSTANT	0.0135	–0.0106	0.0225	–0.0213	–0.0444	–0.0682
	(0.0447)	(0.0460)	(0.0504)	(0.0506)	(0.0918)	(0.0294)
Number of obs.	736	736	523	523	213	213
F (H_0: α_n=0)	13.82	11.80	11.26	11.18	11.56	11.76
F (H_0: μ_i=μ_s)	5.22	5.07	8.70	8.98	2.35	3.55
F (H_0: υ_j=υ_k)	11.69	12.12	8.66	8.90	11.08	12.26
R^2	0.26	0.26	0.20	0.19	0.56	0.56

Notes: Fixed effects panel data estimation, standard error in parentheses.
†, ‡, * denote statistical significance at the 1 per cent, 5 per cent, 10 per cent level respectively.

As regards comparative advantage, *CA*1 and *CA*2 produce statistically significant and positive coefficients in the estimations. This suggests that, controlling for the level of US FDI, the wage rate, and other sector and country-specific effects, a country which has a revealed comparative advantage in the production of a good also attracts US inward processing trade into this sector. This finding is in line with the prediction of international trade theory regarding the dispersion of fragmentation around the globe (Jones and Kierzkowski, 2001). Since the dependent variable and the *CA* variables are measured in percentage terms, the interpretation of the coefficient is straightforward. For the results reported in column 1, an increase in *CA*1 by one percentage point will lead to an increase in US IPT by 0.1 percentage points, *ceteris paribus*.

The *PERFDI* variable turns out to have a positive and statistically significant coefficient. Note that this variable is also measured as a percentage, which makes the interpretation of the coefficient straightforward. For example, in the specification in column 1, an increase in FDI stocks by one percentage point leads to an increase in US IPT of 7.3 percentage points, *ceteris paribus*. The results indicate that the presence of US foreign direct investment in the EU member country has a positive impact on US IPT in that country. This suggests that US multinationals in the EU are engaged in intra-firm trade, they ship intermediate products to their affiliates in the EU. These affiliates process the intermediate goods and re-export them to destinations outside the EU, either back to the parent company in the US, or for final consumption, or further processing to other locations world-wide. In other words, this proportion of production is not intended to serve the European market, which is in contrast with the frequently asserted opinion that US multinationals

in the EU locate there in order to establish a base to serve the European market (Balasubramanyam and Greenaway, 1992; Aristotelous and Fountas, 1996). While our results do, by no means, contradict this assertion, they do suggest, however, that the use of the affiliate in the EU as a base for fragmentation is also an important factor.

The *WAGE* variable is also statistically significant and positive. This implies that higher labour costs seem to make a particular sector in a country more attractive for US inward processing trade; a result that may seem odd at first sight. However, if we think of the *WAGE* variable as a proxy for different skill levels in different sectors, the result does appear reasonable. Higher skilled labour can be expected to be paid higher wage rates than low-skilled labour. The results may, hence, indicate that sectors with higher skill levels receive higher levels of US IPT, *ceteris paribus*.[15]

In order to try to shed some more light on the issue of why a particular EU member state receives US IPT into a particular sector we divide our sample of countries into core and peripheral countries.[16] As Görg and Ruane (1999) find, production activities of US companies show different patterns in EU core and periphery countries. While US companies in core countries are dispersed across all manufacturing sectors, peripheral countries show evidence of sectoral specialisation, US companies are concentrated in a small number of industrial sectors rather than being dispersed across the whole spectrum of manufacturing sectors. Therefore, we may expect that there are different determinants of US IPT in the core and periphery as well. If a large share of IPT is intra-firm trade, IPT will be more sectorally specialised in peripheral countries and may, therefore, be more strongly influenced by the stock of FDI or comparative advantage, *ceteris paribus*.

Columns 3 and 4 in Table 8.3 present the results for the estimation of equation (8.1) using data for core countries. Inspection of the results shows that, in contrast to the results for all EU countries, only the coefficients of the comparative advantage variables are statistically significant. In other words, the presence of US FDI or the level of labour costs in a sector do not seem to explain the location of US inward processing trade across core countries, other things being equal. However, a country's comparative advantage does have a positive effect on its receiving US IPT.

The results in columns 5 and 6 relate to the estimation using data for EU peripheral countries only. It can be seen that, *ceteris paribus*, a country's revealed comparative advantage has a positive effect on the level of US IPT in peripheral countries. The results also show that, for peripheral countries, the presence of US foreign direct investment has a positive impact on the location of US IPT in the country. This suggests that the possibility of fragmented production is at least one of many factors that attract US FDI into a particular sector in an EU peripheral country. However, as columns 3 and 4 show, this does not seem to be an important factor for US IPT in core countries. The results possibly indicate that US multinationals use their affiliates in peripheral countries in part as a base for processing intermediate goods, which are then shipped to destinations outside the EU, either back to the parent company in the US or for final consumption or further processing to other locations world-wide.

The *WAGE* variable is also statistically significant and positive. This implies that higher labour costs, possibly reflecting higher levels of human capital, seem to make a particular sector in a country more attractive for US inward processing trade. The results indicate that sectors with higher skill-levels, higher levels of human capital, receive higher levels of US IPT, *ceteris paribus*.

Conclusions

Fragmentation is referred to as the possibility of splitting up the production process into various components, following, *inter alia*, advances in communications and transportation technologies. This allows plants to outsource intermediate inputs, which enhances trade in components and also fosters the emergence of multinational companies that trade intermediate goods between different plants within the same firm.

There has been a great interest in fragmentation in the recent literature, which examines this phenomenon mainly through the lens of international trade theory, emphasising comparative advantage. The present chapter discusses an alternative approach for looking at the issue of fragmentation, namely an approach based on transaction costs theory and presents an empirical analysis of the determinants of fragmentation.

Our empirical analysis of US inward processing trade in EU countries as a measure of the extent of fragmentation taking place in the EU lends support to the comparative advantage explanation. Countries attract inward processing trade in those sectors in which they have a comparative advantage. Taking this result at face value implies that countries which have a comparative advantage in low (high) skill intensive production will specialise in the production of low (high) skill intensive fragments. One may argue that this raises doubts, from a policy perspective, as to how beneficial fragmentation may be in assisting developing countries in their endeavours to build up a developed competitive industrial structure. However, this is clearly a dynamic question the analysis of which is beyond the scope of the present study.

For countries at the periphery of the EU, the empirical results of our analysis also show that the presence of US foreign direct investment has a positive effect on their receiving US IPT. This suggests that at least part of the inward processing trade going into peripheral countries is intra-firm trade between different plants of the same multinational company rather than trade between different firms. This indicates that it may be worth investigating whether FDI is intended to serve local markets or to shift parts of the production process to locations with lower production costs.

Acknowledgements

The author is grateful to James McCarthy and Adrian Redmond of the Central Statistics Office of Ireland for help with the Eurostat data. Thanks are due to Marius Brülhart, Gerard O'Reilly, Frances Ruane, and participants of the IESG

Easter Conference at Warwick University, April 1999 for helpful comments. The usual disclaimer applies. Financial support from the Arts Faculty Research Fund at University College Cork and the Leverhulme Trust (Programme Grant F114/BF) is gratefully acknowledged.

Notes

1 For example, think of *I* as a desktop computer with *A* and *B* being monitor and CPU respectively. Initially, due to technological requirements or high transportation costs, production cannot be split into the two fragments but a desktop computer has to be produced in its entirety in the production plant.
2 Jones and Kierzkowski (2001) assume that the breaking up of the production process is costless. This assumption is relaxed by, for example, Deardorff (2001).
3 For a discussion of these issues, and empirical evidence, see, for example, Lawrence and Slaughter (1993) and Berman *et al*. (1994) for the US, Feenstra and Hanson (1997) for Mexico, Figini and Görg (1999) for Ireland, and Machin and van Reeden (1998) for a number of OECD countries.
4 These data were available to us on CD ROM: *Eurostat: Intra- and Extra-EU trade (Combined Nomenclature)*.
5 See Baldone *et al*. (2001) for an empirical study of OPT from the EU to Central and Eastern European countries.
6 All the data in this section are for 1994, due to the fact that data constraints on the variables used in the empirical estimation dictate the use of 1994. This implies that all data relate to the EU 12, excluding Austria, Finland and Sweden in 1995.
7 See, for example, Aristotelous and Fountas (1996) and Görg and Ruane (1999, 2000) for further discussions of US investment in the EU.
8 Greece, Ireland, Portugal and Spain are defined as the EU periphery.
9 This section draws on Görg (2000).
10 There is a possible problem with the model in equation (8.1) in that *PERFDI* may itself be positively influenced by a country's comparative advantage (Milner and Pentecost, 1996) and the level of labour costs (Barrell and Pain, 1996). We therefore regress *PERFDI* on *CA* and *WAGE* in an auxiliary regression and use the residual from this regression as a proxy for US FDI. In other words, the residual reflects the share of US foreign direct investment that is not explained by either comparative advantage or labour costs and is, therefore, employed as a proxy for *PERFDI* in equation (8.1).
11 The empirical measurement of comparative advantage is problematic, and this fact has been discussed extensively in the literature, see Deardorff (1980), Ballance *et al*. (1987).
12 One may argue that the *CA*1 and *CA*2 variables are correlated with the *WAGE* variable. In a correlation analysis we found only a very weak correlation (correlation coefficient equal to 0.35 between *CA*1 and *WAGE* and 0.05 between *CA*2 and *WAGE*). The correlation coefficient between *CA*1 and *CA*2 is 0.63.
13 We use lagged variables for all right-hand side variables. In the case of the revealed comparative advantage measures, this is partly because of the possible endogeneity between *CA*1 or *CA*2 and the dependent variable, if data for the same year were used.
14 We do not present the empirical results for the estimation of the auxiliary regression herein. They can be obtained from the author upon request. The results suggest that comparative advantages have positive and labour costs have negative effects on the level of US FDI.
15 As a robustness check we also estimated equation (8.1) excluding sector and country dummies. We find that the R^2 increases considerably when including these dummies. This may, on the one hand, reflect that production in some sectors may be more likely to

be fragmented than in others. For example, production of PCs may easily be fragmented whereas the production of wooden furniture can be assumed to be less likely to be split up into various components. On the other hand, some countries may offer specific advantages which are not picked up in our model, such as cultural factors or language.

16 The latter sample consists of data for Greece, Ireland, Portugal and Spain while the former contains data for the other eight member states.

References

Aristotelous, K. and S. Fountas (1996) 'An empirical analysis of inward foreign direct investment flows in the EU with emphasis on the market enlargement hypothesis', *Journal of Common Market Studies*, 34: 571–83.

Arndt, S.W. (1997) 'Globalization and the gains from trade', in Jaeger, K. and K.J. Koch (eds) *Trade, Growth, and Economic Policy in Open Economies*, New York: Springer Verlag.

Balasubramanyam, V.N. and D. Greenaway (1992) 'Economic integration and foreign direct investment: Japanese investment in the EC', *Journal of Common Market Studies*, 30: 175–93.

Baldone, S., F. Sdogati and L. Tajoli (2001) 'Patterns and determinants of international fragmentation of production: evidence from outward processing trade between the EU and Central Eastern European countries', *Weltwirtschaftliches Archiv*, 137: 80–104.

Ballance, R.H., H. Forstner and T. Murray (1987) 'Consistency tests of alternative measures of comparative advantage', *Review of Economics and Statistics*, 69: 157–61.

Baltagi, B.H. (1995) *Econometric Analysis of Panel Data*, Chichester: John Wiley and Sons.

Barrell, R. and N. Pain (1996) 'An econometric analysis of US foreign direct investment', *Review of Economics and Statistics*, 78: 200–7.

Berman, E., J. Bound and Z. Griliches (1994) 'Changes in the demand for skilled labor within U.S. manufacturing: evidence from the *Annual Survey of Manufacturers*', *Quarterly Journal of Economics*, 109: 367–97.

Caves, R.E. (1996) *Multinational Enterprise and Economic Analysis*, second edition, Cambridge: Cambridge University Press.

Coase, R.H. (1937) 'The nature of the firm', *Economica*, 4: 386–405.

Coase, R.H. (1992) 'The institutional structure of production', *American Economic Review*, 82: 713–719.

Deardorff, A.V. (1980) 'The general validity of the law of comparative advantage', *Journal of Political Economy*, 88: 941–57.

Deardorff, A.V. (1998) *Fragmentation in Simple Trade Models*, Research Seminar in International Economics Working Paper n o. 422. University of Michigan.

Deardorff, A.V. (2001) 'Fragmentation across cones', in S. Arndt and H. Kierzkowski (eds) *Fragmentation: New Production Patterns in the World Economy*, Oxford: Oxford University Press.

Driffield, N. and M. Munday (2000) 'Industrial performance, agglomeration, and foreign manufacturing investment in the UK', *Journal of International Business Studies*, 31 (1): 21–37.

Dunning, J.H. (1997) 'The European Internal Market Programme and inbound foreign direct investment', *Journal of Common Market Studies*, 35: 1–30.

External and Intra-EU trade database, Luxembourg: Eurostat, European Commission.

Feenstra, R. (1998) 'Integration of trade and disintegration of production in the global economy', *Journal of Economic Perspectives*, 12: 31–50.

Feenstra, R.C. and G.H. Hanson (1997) 'Foreign direct investment and relative wages: evidence from Mexico's maquiladoras', *Journal of International Economics*, 42: 3–4.

Figini, P. and H. Görg (1999) 'Multinational companies and wage inequality in the host country: the case of Ireland', *Weltwirtschaftliches Archiv*, 135: 594–612.

Görg, H. (2000) 'Fragmentation and trade: US inward processing trade in the EU', *Weltwirtschaftliches Archiv*, 136: 403–22.
Görg, H. and F. Ruane (1999) 'US investment in EU member countries: the internal market and sectoral specialization', *Journal of Common Market Studies*, 37: 333–48.
Görg, H. and F. Ruane (2000) 'European integration and peripherality: lessons from the Irish Experience', *The World Economy*, 23: 405–21.
Greenaway, D. and C. Milner (1986) *The Economics of Intra-Industry Trade*, Oxford: Basil Blackwell.
Greenaway, D. and J. Torstensson (2000) 'Economic geography, comparative advantage and trade within industries: evidence from the OECD', *Journal of Economic Integration*, 15: 260–80.
Harris, R.G. (1995) 'Trade and communication costs', *Canadian Journal of Economics*, 28: 46–75.
Harris, R.G. (2001) 'A communication based model of global production fragmentation', in S. Arndt and H. Kierzkowski (eds) *Fragmentation: New Production Patterns in the World Economy*, Oxford: Oxford University Press.
Helleiner, G. (1981) *Intra-Firm Trade and the Developing Countries*, London: Macmillan.
Jones, R.W. and H. Kierzkowski (1990) 'The role of services in production and international trade: a theoretical framework', in R.W. Jones, and A.O. Krueger (eds) *The Political Economy of International Trade*, Oxford: Basil Blackwell.
Jones, R.W. and H. Kierzkowski (2001) 'A framework for fragmentation', in S. Arndt and H. Kierzkowski (eds) *Fragmentation: New Production Patterns in the World Economy*. Oxford: Oxford University Press.
Kierzkowski, H. (2001) 'Joining the global economy: experience and prospects of the transition economies', in S. Arndt and H. Kierzkowski (eds) *Fragmentation: New Production Patterns in the World Economy*, Oxford: Oxford University Press.
Lawrence, R.Z. and M. Slaughter (1993) 'International trade and American wages in the 1980s: giant sucking sound or small hiccup?', *Brookings Papers on Economic Activity: Microeconomics*, 2: 161–226.
Liu, P.-W. and X. Yang (2000) 'The theory of irrelevance of the size of the firm', *Journal of Economic Behavior and Organization*, 42: 145–65.
Lundbäck, E.J. and J. Torstensson (1998) 'Demand, comparative advantage and economic geography in international trade: evidence from the OECD', *Weltwirtschaftliches Archiv*, 134: 230–49.
Lyons, B.R. (1995) 'Specific investment, economies of scale, and the make-or-buy decision: a test of transaction cost theory', *Journal of Economic Behavior and Organization*, 26: 431–43.
Machin, S. and J. Van Reenen (1998) 'Technology and changes in skill structure: evidence from seven OECD countries', *Quarterly Journal of Economics*, 113: 1215–44.
Milner, C. and E. Pentecost (1996) 'Locational advantage and US foreign direct investment in UK manufacturing', *Applied Economics*, 28: 605–15.
Neven, D.J. (1990) 'EEC integration towards 1992: some distributional aspects', *Economic Policy*, 10: 14–62.
Ruane, F. and H. Görg (2001) 'Globalisation and fragmentation: evidence for the electronics industry in Ireland', in S. Arndt and H. Kierzkowski (eds) *Fragmentation: New Production Patterns in the World Economy*, Oxford: Oxford University Press.
Svensson, R. (1996) 'Effects of overseas production on home country exports: evidence based on Swedish multinationals', *Weltwirtschaftliches Archiv*, 132: 304–29.
UNCTAD (1997) *World Investment Report 1997: Transnational Corporations, Market Structure and Competition Policy*, New York: United Nations.
Williamson, O.E. (1975) *Markets and Hierarchies: Analysis and Antitrust Implications: A Study in the Economics of Internal Organizations*, New York: Free Press.
Williamson, O.E. (1979) 'Transaction-cost economics: the governance of contractual relations', *Journal of Law and Economics*, 22: 233–61.

Yeats, A.J. (2001) 'Just how big is global production sharing', in S. Arndt and H. Kierzkowski (eds) *Fragmentation: New Production Patterns in the World Economy*, Oxford: Oxford University Press.

9 Technological innovation

The prime cause of international trade

Jitendralal Borkakoti

The objective of this chapter is to focus on the implicit role of technological innovations in facilitating and generating trade among the nations of the world. Although it was obvious that the invention of steam engine, textile machines etc. led to rapid economic growth in the UK with the consequent expansion of international trade, technological innovation was not at the core of theoretical paradigms in classical economics. The Ricardian model, which perfunctorily captures international differences in technology, was followed by the Heckscher–Ohlin–Samuelson (HOS) model which abstracts from technological differences altogether and which dominated trade theory from the 1950s until the early 1980s. It will be argued that in reality technological innovation plays a tacit and unrecognised role here too. It was the advent of the neo-technology theory that provided a hiatus, after which the current models of endogenous innovation and international trade begin to give technological innovation the prime role that it deserved all along. We begin with a discussion of the various forms in which technological innovations may take place in the first section. We then proceed to present a historical sketch of international trade in the second section. The third section critically examines how classical economists emphasised capital accumulation, but neglected the important role of technological innovation. In the fourth and fifth sections, we argue how technological innovation silently, but in an empirically relevant way, plays an important role in the earlier trade models of the 1960s and 1970s, and discuss how the more recent new trade models put technological innovation at the heart of economic growth and international trade.

Forms of technological innovation

It seems instructive to consider the various forms of technological innovation which, by definition, imply that something new has been invented and developed. There are five forms of technological innovations, namely, process innovation, product innovation, material-change technological innovation, epochal innovation, and lastly, general autonomous technological innovation. These categories are not necessarily independent of each other. For example, the invention of a new product

is accompanied by a new process of production; or product innovation may occur as a result of a change in the input mix with which the product is made.

It is epochal innovation that needs some explanation. An innovation is epochal if its effect ripples through many economic activities over a significant period of time (several decades or even a century). The invention of electricity, for example, led to many process innovations (capital goods or machines powered by electricity) and also to many product innovations (mainly durable consumption goods). Over a period of four to five decades, electric power was used increasingly to run machines in many industries. Also, over a similar epoch, many consumption goods were invented which use electricity, ranging from electric fans to electric toothbrush. Another example of epochal innovation is the steam engine which, over a period of time, was used not only to sail ships but also to produce textiles and yarns. The current epochal innovation is the 'micro chip' technology which has already led to many new and improved products.

Lastly, the general autonomous technological innovations may mean either of the following. First, it may be defined in terms of 'technological spread'. With regards to electricity, say, one may argue that an economy continues to have more miles of electric trains, or more streets have electric lights or the number of households possessing washing machines increases, or the number of offices and factories using computers steadily increases. Secondly, it may be defined in terms of 'learning by doing' (a function of the cumulative output), that is, the economy learns to do things better over time which gets reflected in steady increases in productivity.

A historical approach

If one analyses the origins of international trade, one would find three important stylised facts: (1) the discovery of new products which led to international trade; (2) innovations in shipping technology that made it possible to transport heavy goods from one country or continent to another; and (3) technological innovations in production and transport during the industrial revolution.

International trade developed in classical antiquity. Archaeological discoveries suggest that nations or jurisdictions traded with one another from at least the third millennium BC. The Phoenicians in Tyre (Lebanon) and Carthage (Tunisia) were engaged in trade spanning three continents (Warmington, 1969). Around 100 BC trade between Rome and China via the Silk Route involved the exchange of high-value commodities, such as Chinese silk, Roman wool, and precious metals. Some of these goods were invented and developed over the centuries prior to 100 BC and some goods (like precious metals) came into existence purely through discovery and invention of new processes. During the fifteenth and sixteenth century, because of improved technology of ocean-going ships, trade expanded between the Orient and Europe. During the seventeenth century, the discovery of America led to trade in new goods like tobacco, while Spain found gold and silver deposits in South America, and Europe imported luxury goods in exchange for precious metals. In the mid-eighteenth century, the industrial revolution propelled entirely by new

innovations (steam engines, textile machines, railways, ships etc.) expanded international trade (a five-fold increase between 1750 and 1914) once again. A definite pattern of trade emerged in the form of Europe exporting manufactured goods and importing raw materials. Europe also imported grain from North America, Australia, Argentina and India. The invention of the steam-powered railway engine led to dramatic improvement in European trade internally.

Ships have played an important historical role in expanding international trade. Phoenicia as early as in 2000 BC constructed ships capable of carrying large cargo, and the Phoenicians travelled around the Mediterranean, West Africa, and also to Britain for trade in tin. The Greeks, the Romans, and the Vikings improved shipping technology over the centuries. China also developed her shipping industry by the ninth century, and the Chinese junks were carrying goods to Indonesia and India, and by the fifteenth century, to east Africa. Shipping technology also improved in medieval Europe, and a clear distinction was made between warships and merchant ships. During the seventeenth century, Portugal and Spain developed caravels (capable of carrying 110 tonnes of cargo) with three or four masts, and the foremast carried square sails and the rest carried lateen sails. Christopher Columbus sailed in a caravel and discovered America in 1492. European ships became larger, and warships capable of carrying 100 guns were built. The British shipbuilding industry developed fast, and Nelson's HMS Victory was built in the middle of the eighteenth century. In the meantime, America developed clipper ships (long, slender and sharp-bowed) which were faster. This technological improvement greatly increased a ship's capability for long-distance trade between United States and China, Britain and India, and United States and Britain. Again, the United States led the world in building the first steam-powered ship, which crossed the Atlantic in 1819 (from Savannah in Georgia to Liverpool). The technology of steam engines improved in 1854 (double-expansion engine), and also in 1873 (triple-expansion engine); ship design also changed from paddle technology to turbine technology. Then in the twentieth century, shipping technology continued to improve with diesel-powered ships, ultimately leading to nuclear-powered ships in the 1960s. On commercial shipping two other important developments were: (a) container ships; and (b) tankers. There is no doubt that the expansion of world trade is closely related to the development of the world transport systems, and the contribution of the development in shipping technology is central to the growth of trade.

The third fact is that the history of the industrial revolution is fundamentally the history of technological innovation. One only needs to glance at the history of textiles, iron and steel, the steam engine, and the railways. The invention of the 'flying shuttle' by John Kay in 1733 speeded up weaving. James Hargreaves in 1765 invented the 'spinning jenny' – a machine that could spin six threads of yarn at once. Richard Arkwright's 1769 invention of the 'water frame' – a spinning machine powered by a water wheel was improved by Samuel Crompton in 1779. Edmund Cartwright invented a power loom in 1785, using a steam-powered machine. Textile production began to take place in factories on a large scale. With regard to iron production, Britain had a relatively small iron industry in 1750. A

new method called the puddling and rolling method of turning pig iron to wrought iron quickly and cheaply was in use by 1784. In 1828, James Nielsen reduced the quantity of fuel by pumping hot air into the furnaces, and this lowered cost. The iron making industry grew very fast and iron production, which stood at about 30,000 tons in 1750, increased to six million tons in 1870. Production of steel also grew, especially after Henry Bessemer's invention of a 'converter' in 1856, and William Siemens's invention of the 'open hearth' method in 1867. British steel production grew from 60,000 tons in 1800 to two million tons in 1880. Note that the invention of steam engine – a technology adapted for various industrial activities – was an epochal innovation and was at the core of industrial revolution. Thomas Savery's steam engine (invented in 1698) pumped out water from tin mines, and Thomas Newcomen improved this machine in 1712. But these machines were not capable of driving rotary motion. It was James Watt (helped by Matthew Bolton) who in 1788 produced a steam engine which used the piston and the cylinder, driving rotary motion. This technological innovation enabled many industries to use steam engines as a source of power – leading to large-scale production. One important application of the steam-engine technology was in the railways. Richard Trevithick developed the railway engine – a steam-powered locomotive which could pull several carriages along rails. Steam locomotives became popular, and Britain had 22,000 miles of railway track in 1900. The benefits of steam-powered machines also accrued to agriculture: steam-powered reapers and threshers were used from 1850, steam engines pumped logged water from land areas which could then be cultivated.

Classical economists

It is fair to say that none of the classical economists put technological innovation at the heart of economic growth and/or international trade. In fact, the terms 'technological innovation' or 'technical change' do not appear in their work; they refer only to machines. Adam Smith experienced the first stirrings of industrial revolution, and history reveals that spinning jenny and 'water frame' textile machines were invented while he was working on his *Wealth of Nations* (1776). David Ricardo, James Mill, Robert Torrens, Robert Malthus, Friedrich List, John Mill and William Thornton virtually lived through the industrial revolution as it relentlessly progressed.

Smith's emphasis on 'vent for surplus', as an engine of trade-driven growth, reflected new products (tea, tobacco, spices) and cheaper products (cotton, wheat) imported from the colonies. But this phenomenon was a consequence of various technological innovations. In his *Principles*, Ricardo (1817) has a chapter on machinery (chapter 31) where he discusses the impact of machinery on different classes (namely, landlords, capitalists and workers). In 1820, McCulloch stated in *Edinburgh Review* that machines would lower wage rates; his argument hinged on a distinction between fixed capital (machines) and circulating capital such that circulating capital diminished as fixed capital increased. Since employment and the wage rate depended on circulating capital, wages would fall and workers would

be hurt as a consequence of diminished circulating capital. On the other hand, in a letter to McCulloch (cited in Sraffa, 1970), Ricardo argued that use of machines never diminished demand for labour and it never caused a fall in wages. However, this robust view changed dramatically in his *Principles* (third edition) and painstakingly using a numerical example, Ricardo came to the conclusion that the employment of machinery is 'frequently detrimental' to the interests of the labouring class. In his chapter on foreign trade (chapter 7), Ricardo discusses how an invention of a new process of wine-making increases productivity in wine production in England, and how this redistributes 'specie' and increases both quantities of commodities and 'general prices'. However, Ricardo did not explicitly relate process improvement to labour productivity in his famous example of comparative advantage where labour productivity (or labour requirement per unit of output) is assumed as given. But, a link between exports and improved machinery is clearly mentioned, although not pursued rigorously. The argument is that prices depend on cost of production and the cost of production is reduced by 'employing improved machinery', and hence these commodities with lower prices can be sold in foreign countries.

Malthus (1836) was struck by the invention of textile machinery in Britain, and took inventions of machines ('the result of ingenuity of man') as substituting 'manual exertions'. In his *Principles* (second edition, 1836), he said:

> When a machine is invented, which, by saving labour, will bring goods into the market at a much cheaper rate than before, the most usual effect is such an extension of the demand for the commodity, by its being brought within the power of a much greater number of purchasers, that the value of whole mass of goods made by the new machinery greatly exceeds their former value; and, notwithstanding the saving of labour, more hands, instead of fewer, are required in the manufacture (p. 352).

Thus, Malthus not only believed in the ability of new inventions to increase output and consumption, but also in the employment enhancement impact of machines, provided that it leads to 'opening of a sufficient market and an adequate increase in consumption' (p. 360). Malthus's conclusion is that three factors cause production to increase and these are: accumulation of capital, fertility of soil, and inventions to save labour. It is clear that Malthus was the only exception among the classical economists who did understand the nature of innovation.

Standard trade models

Let us consider the two standard trade models, the Ricardian model and the Heckscher–Ohlin–Samuelson (HOS) model. The former underlines differential labour productivity and the latter highlights differential factor endowments as causes of international trade. Ricardo did not discuss the causes of international differences in labour productivity, but the usual interpretation of the fixed coefficient production function is that it represents international technological difference.

However, such an interpretation of labour productivity clashes with the relative abundance of complementary factors of production (like land in corn production). Attempts have been made to adapt Ricardo's single factor model to the multi-factor real world, as in Bhagwati (1964). As we have seen in the previous section, Ricardo himself did not link invention of machines to comparative advantage.

The HOS theorem – arguably the most dominant paradigm in trade theory – emphasises relative abundance of factors as the prime cause of trade (see Borkakoti, 1998). Countries use identical technologies to produce similar goods, and export the goods that intensively utilise the factor of production more abundant in that country. But empirical tests have found the HOS theory to be empirically unacceptable particularly in its most robust $2 \times 2 \times 2$ form where two generic factors of production, physical capital and labour are deemed to exist. In the so-called three-factor tests, human capital stands out as the most crucial variable explaining trade. But, this is an embarrassing piece of empirical evidence in so far as human capital is fundamentally related to new technology and new skills. Human capital is simply *knowledge* incorporated in human beings through investment in education and training.

As far as physical capital is concerned, note first that this variable is usually measured by the sum of past investments in a particular sector, with linear depreciation. Now, it is of crucial importance to realise that manufacturing investment itself is affected by innovation. During the last fifty years, manufacturing investment has largely taken place to produce newly-invented goods. In the 1950s and 1960s, a lot of new investment had gone into producing consumer durables (radios, televisions, washing machines, dishwashers, cars and so on), in the 1980s new investment went into electronic goods (camcorders, CD players, mobile phones, computers, etc.). If the hypothesis that manufacturing investment is led by new innovations bears some empirical truth, then it makes little sense to take capital as a given generic factor of production, and to argue that relative factor endowments will determine the pattern of trade. This is because the process of capital accumulation is related to innovation in an intrinsic manner.

There is a more fundamental problem with the HOS model. This is an issue raised in a classic paper by Oniki and Uzawa (1965) – a paper largely ignored by Anglo-Saxon economists (only Harry Johnson paid tribute to this paper). This paper presents a dynamic model with two countries where endogenous capital accumulation takes place and labour grows exogenously, and demonstrates that a globally stable steady-state exists. The proposition that emerges (see Borkakoti, 1998: 289) is the following:

> If tastes and technology are internationally identical, capital accumulation takes place endogenously, and population grows exogenously at an internationally identical rate, then at the steady-state equilibrium, both countries have an identical steady-state level of capital–labour ratio, and international trade vanishes.

However, it is easily shown that if tastes are not identical internationally (average saving propensities are different), then trade exists at the steady-state such that the

country with the higher propensity to save exports the relatively capital-intensive consumption good. In an extension of this model, Borkakoti (1995) shows that steady-state trade exists when the country experiencing Hicks-neutral technological progress in the consumption goods sector has a relatively higher steady-state capital–labour ratio, and exports the consumption good to her trade partner who does not experience technological change. The cause of trade in this instance is process innovation in one country. Thus, these dynamic models reveal that, for endowment-based trade to continue at the steady-state, one needs to assume international differences either in technology or in the patterns of demand. This theoretical finding is supported by empirical evidence in Trefler (1993, 1995). Trefler (1993) points out that two central tenets of the traditional HOS theory are at odds with real world experience. The first is the factor price equalisation theorem, a process whereby international trade equalises the returns to factors of production across countries. On the contrary, international wage data suggests differences across countries. The second assumption of HOS theory is identical production technologies for the same good across nations, which is patently unrealistic. Trefler (1995) puts forward two interesting notions, namely, (1) 'the case of missing trade'; and (2) 'the endowment paradox'. Missing trade implies that factor endowment differences do not explain trade patterns across nations. Plotting the abundant factors against each country, he finds that 'the rich countries tend to be scarce in most factors and the poor countries tend to be abundant in all factors'. Trefler calls this finding 'the endowment paradox'. Having found poor empirical support for the HOS model (thirty-three countries, nine factors of production, using 1983 data), Trefler decided to investigate why the empirical results were so unsupportive of the HOS theory. The explanations lie in terms of: (a) technology differences; and (b) consumption or demand differences. These empirical results coincide with the steady-state predictions of the Oniki–Uzawa type dynamic models, because steady-state trade is zero unless technological differences are introduced or differential time preference is assumed across nations.

Now one begins to wonder why such a theoretical hysteresis persists and why economists desist from overthrowing a theory (HOS) based on one conjecture and replacing it by another theory based on a separate conjecture. Why, for example, do we not assume the exact converse of the HOS theory: that all countries are identical in factor endowments, but the state of technology is internationally different such that international trade is based on technological differences? From this theory would have emerged the theorem of non-equalisation of factor prices, and we would have felt more comfortable as theoretical predictions would not be so blatantly at odds with the facts of the real world. The only explanation at hand seems to be our reluctance to radically change our earlier conjecture – a case of theoretical hysteresis. Also, methodologically, such hysteresis may be considered irrational as we seek to explain real world phenomena with models containing postulates that contradict the facts of the real world.

Modern trade models

The era of modern trade theories began with the advent of the neo-technology and neo-factor proportions hypothesis – triggered by the Leontief paradox (1953). This paradox stated that the pattern of world trade was at variance with the postulates of the HOS theory, the capital-rich USA, for example, exported labour-intensive goods (or intensive in skilled labour embodying technology). The neo-factor proportions hypothesis, which asserts that trade patterns are determined by factors theoretically and empirically more relevant than physical capital and raw labour, encompasses the human-capital hypothesis, the skill-intensity hypothesis, the research and development (R&D) hypothesis, and the scale-economies hypothesis. The results of the empirical tests of these hypotheses are very strong indeed, although these are essentially piecemeal empirical hypotheses.

The neo-technology theory, first mooted by Posner (1961), asserts that technological progress in terms of both product innovation and process innovation is the prime cause of trade. The trade advantage (comparative if caused by process innovation, but absolute if caused by product innovation) which a country acquires by virtue of new innovations is temporary, because after a lag in time, the rest of the world imitates the new technology. World trade thrives because of a continuous process of creation and diffusion of new technology. The product-cycle hypothesis, which emphasises certain aspects at the microeconomic level, falls within the class of the neo-technology theory. Krugman (1979a, 1979b) lucidly brings out the logic of trade based on product innovation and process innovation by using elegant mathematical models. The empirical results, (see, for example, Hufbauer (1966, 1976), Soete (1981), Wells (1969)), do provide overwhelming evidence that technological innovation is a major determinant of trade performance. It should be pointed out that some economists consider the empirical specifications of the variables, which capture the basic tenets of the neo-factor proportions and neo-technology hypotheses, as different facets of the same phenomenon, while some other economists distinguish between the two sets of variables as emanating from two entirely different phenomena. It is generally accepted that physical capital, human capital, skilled labour, wages per worker, measures of materials or natural resources, and the consumer goods ratio are the variables which capture the neo-factor proportions theory. While R&D acitivity, first-trade date, patents, scale economies, industry concentration, and a measure of product differentiation are the variables which capture the neo-technology theory. See, for example, Wolter (1977), Hirsch (1976), and Goodman and Ceyhun (1976).

Turning now to intra-industry trade models, we note that the development of these models in the late 1970s and 1980s was a consequence of a determined effort to replace the assumption of perfect competition in product markets by monopolistic competition and oligopoly. The dynamic questions of technological innovation raised by neo-technology theory are not considered in this class of models. A range of models attempts to explain intra-industry trade both in horizontally differentiated products, as for example, Krugman (1979b) with love for variety and Lancaster (1980) with a preference for a variety, and in vertically differentiated

products, Falvey (1981). The models emphasise demand specifications with the Dixit–Stiglitz preference functions for the models with horizontal product differentiation when all varieties are symmetrically consumed, and with the Hotelling–Lancaster preference function when only a preferred variety is consumed. Technological innovation is neutralised by assuming a simple production structure, often with a single factor while the phenomenon of increasing returns to scale is incorporated by introducing a fixed-cost component. The models dealing with vertical product differentiation reintroduces the notion of factor endowments since a high quality product is indexed by the requirement of a relatively high capital–labour ratio. Although these models assume product differentiation as given, it is true to say that technological innovation leads to product differentiation through R&D investment. In an industry where a large number of horizontally differentiated varieties are produced, each firm tends to have a degree of monopoly for its product as opposed to the absolute, but temporary, monopoly of an industry in the neo-technology theory. When it comes to vertical product differentiation, the role of technological innovation seems to be apparent, and the degree of monopoly in an industry (producing a continuum of vertically differentiated products) becomes stronger as higher quality goods are innovated and produced. This chain of logic reveals a clear proximity to the neo-technology theory, as technological innovation plays an important role. This proximity has been recognised in the 1990s leading to a host of models dealing with endogenous growth and international trade.

The term 'endogenous growth' owes its genesis to endogenous knowledge creation either because of learning by doing, or because of deliberate efforts to invent new products and processes by investing in R&D. The main focus of the models is on how intra-industry trade is dynamically generated by endogenous innovation within a general equilibrium framework. For example, the Grossman–Helpman model (1991, chapter 7), analyses how the pattern of trade evolves over time as new varieties of differentiated products are generated by investing in R&D activities. On the other hand, the DOS model, Dinopoulos, Oehmke and Segerstrom (1993), analyses how firms engage in R&D races and the winners gain temporary monopoly in the production of the newly innovated goods for a defined period of the patent monopoly. The existence of a steady-state is proved, and it is also shown how product-cycles may take place. These dynamic models treat endogenous technological innovation as the central cause of international trade, and one immediately realises the proximity of the conclusions of these models with those of the neo-technology theory.

Conclusions

Our proposition is that the prime cause of international trade has always been technological innovation. In early periods of history technological progress in shipping led to expansion of trade, and newly discovered goods in the new lands became traded goods. Also, the history of the industrial revolution is nothing but a history of technological innovations – leading to mass production of many goods and many-fold expansions of world trade. The classical economists did not

appreciate the economic implications of technological innovation; although many of them lived through the industrial revolution, they were mainly concerned with the employment implications of the 'machine' – a term used for technological innovation.

Although the Ricardian theory is based on international difference in labour productivity, its production-function interpretation admits technological differences. The HOS factor-endowment model in the form of two generic factors, capital and labour, appears silent about technological innovation by assuming internationally identical technology, but we argue that technological innovations have an important role in the process of both human and physical capital formation. The neo-technology theory clearly hypothesises that technological innovation dynamically generates world trade through a process of innovation and imitation. The static intra-industry trade models, although seemingly oblivious about technological innovations, do assume product innovation implicitly as product differentiation (especially the vertical type) is generated by R&D activities in the real world. This fact is clearly recognised in the models of endogenous growth and international trade. All empirical studies in trade and technology clearly show that the independent variables, which directly and indirectly capture technological innovation, are always highly statistically significant. Thus on historical, theoretical and empirical grounds, we argue that technological innovation is the prime factor that drives the wheels of world trade. This also has policy implications. Countries wishing to export more goods have to move up the technological ladder through greater investment in knowledge embodying human capital.

Acknowledgement

I wish to thank Ephraim Kebede for research assistance.

References

Bhagwati, J. (1964) 'The pure theory of international trade: a survey', *Economic Journal*, 74: 1711–84.

Borkakoti, J. (1995) *Technological Progress and Capital Accumulation Together in a Dynamic Model of International Trade*, Economics Discussion Papers no. 15, December, Middlesex University Business School.

Borkakoti, J. (1998) *International Trade: Causes and Consequences*, London: Macmillan.

Dinopoulos, E., Oehmke, J.F., and Segerstrom, P.S. (1993) 'High-technology industry trade and investment: the role of factor endowments', *Journal of International Economics*, 34: 49–71.

Falvey, R.E. (1981) 'Commercial policy and intra-industry trade', Journal of *International Economics*, 11: 495–511.

Goodman, B. and Ceyhun, R. (1976) 'US export performance in manufacturing industries: an empirical investigation', *Weltwirtschaftliches Archiv*, 112: 525–55.

Grossman, G.M. and Helpman, E. (1991) *Innovation and Growth in the Global Economy*, Cambridge MA: MIT Press.

Hirsch, S. (1974) 'Capital or technology? Confronting the neo-factor proportions and neo-technology accounts of international trade', *Weltwirtschaftliches Archiv*, 110: 535–63.

Hufbauer, G.C. (1966) *Synthetic Materials and the Theory of International Trade*, London: Gerald Duckworth.
Hufbauer, G.C. (1970) 'The impact of national characteristics and technology on the commodity composition of trade in manufactured goods', in R. Vernon (ed.) *The Technology Factor in International Trade*, New York: NBER.
Krugman, P. (1979a) 'A model of innovation, technology transfer and the world distribution of income', *Journal of Political Economy*, 87: 253–66.
Krugman, P. (1979b) 'Increasing returns, monopolistic competition and international trade', *Journal of International Economics*, 9: 469–79.
Lancaster, K. (1980) 'Intra-industry trade under perfect monopolistic competition', *Journal of International Economics*, 10: 151–76.
Leontief, W.W. (1953) 'Domestic production and foreign trade: the American capital position re-examined', *Proceedings of the American Philosophical Society*, 97(4): 332–49.
Malthus, T.R. (1836) *Principles of Political Economy*, second edition, in E.A. Wrigley and D. Souden (eds) (1986) *The Works of Thomas Robert Malthus*, London: Pickering and Chatto.
Oniki, H. and Uzawa, H. (1965) 'Patterns of trade and investment in a dynamic model of international trade', *Review of Economic Studies*, 32: 15–37.
Posner, M.V. (1961) 'International trade and technical change', *Oxford Economic Papers*, 13: 321–41.
Ricardo, D. ([1817] 1996) *Principles of Political Economy and Taxation*, third edition, Amherst: Prometheus Books.
Soete, L.L.G. (1981) 'A general test of technological gap trade theory', *Weltwirtschaftliches Archiv*, 117: 638–60.
Smith, A. ([1776] 1994) *An Enquiry into the Nature and Causes of the Wealth of Nations*, New York: Random House.
Sraffa, P. (1970) *The Works and Correspondence of David Ricardo*, Cambridge: Cambridge University Press.
Trefler, D. (1993) 'International factor-price differences: Leontief was right!', *Journal of Political Economy*, 101: 961–87.
Trefler, D. (1995) 'The case of missing trade and other mysteries', *American Economic Review*, 85: 1029–46.
Warmington, B.H. (1969) *Carthage*, reprinted 1993, New York: Barnes and Noble.
Wells, L.T. (1969) 'Test of a product cycle model of international trade: US exports of consumer durables', *Quarterly Journal of Economics*, 83: 152–62.
Wolter, F. (1977) 'Factor proportions, technology and West German industry's international trade patterns', *Weltwirtschaftliches Archiv*, 113: 250–67.

10 NAFTA, environmental regulations and firm strategies

Alan M. Rugman and John Kirton

Traditionally, environmental regulatory barriers have posed a specific set of threats that restrict the strategies of firms engaged in international business. The classic threat is from foreign environmental regulations that deny access to the large, lucrative export markets. Such regulatory barriers have been particularly formidable when they moved to higher levels, were backed by powerful coalitions of protectionist industries and environmental groups in the foreign market, and administered by a trade dispute system in large national governments over which outside firms from smaller countries had little control. In such situations the time and expense of litigation and lobbying, even with the full support of one's home government, could be an enormous competitive disadvantage for a firm. The major alternative response, available primarily to those large firms with vast resources and long time horizons, was to produce at home to meet the stringent regulations in the large export market (Vogel, 1995), calculating that these high and ever rising regulations would keep one's competitors, foreign and domestic, at bay.

Today, however, firms face a much more complex situation. Environmental regulations are proliferating at the local, national and international levels. They are expanding from product to production/processing and distribution/disposal phases, and intensifying conflict among industries in different sectors. At the same time, the advent of internationally-integrated production systems is making such local and national regulatory borders increasingly costly, as firms build a larger base enabling them to compete on a fully global scale. This is true even as the rise of multinational production and international business alliances allows firms more readily to produce and exert influence within once closed foreign markets. Finally, to help manage these new intersections of opening markets and compounding environmental regulations, there has arisen a new array of trade liberalisation agreements, the first of which is NAFTA, with new rules for trade-environment integration and new institutions to ensure that the values of both environmental protectors and trade liberalisers are simultaneously enhanced.

This more complex regulatory and competitive environment, and the array of international institutions which govern it, presents firms with new obstacles and opportunities in their response to business challenges abroad. This chapter out-

lines the expanded array of corporate and political strategies which firms now have available in this complex institutional environment, and identifies how firms at different stages of internationalisation are best able to benefit from particular instruments within this menu. This chapter thus begins by identifying the key factors that have changed the regulatory, business, and international institutional environment facing firms – the new conditions of complex institutional responsiveness. It then discusses the new corporate strategies firms now have available, and the new repertoire of political strategies that the new institutional complexity opens for them. It concludes by analysing which of these new corporate and political strategies are likely to be most appealing to, and hence adopted by, firms, from domestic producers through to transnational firms at different levels of internationalisation.

Complex institutional responsiveness

Traditionally, all but the largest firms operating as monopolies or oligopolies in relatively closed markets have needed to be responsive to challenges from competing firms and their home government's abroad. But now, virtually all firms have been forced to respond to such international competition. The widespread demise of closed regimes, along with major multilateral and regional trade liberalisation during the 1990s, has opened up a world market. In seeking to enhance their competitiveness in this far more international arena, firms face new complexities in regulatory, production and international institutional conditions.

The first set of complexities arises from behind-the-border defences of environmental regulations and standards, which the reduction of border tariffs and quotas have rendered more visible and valuable as protectionist devices. The recent era of rising environmental consciousness has added new complexity to the traditional environmental regulatory barriers firms face in foreign markets, and at home as well. Mass public concern with environmental protection has fuelled ever more stringent and rapidly changing regulations, with more flexible enforcement schedules, trade-offs between target levels and deadlines, compliance assistance for firms, and, under the precautionary principle, more flexible standards for scientific proof. Subnational governments have begun to adopt a rapidly changing array of local regulations, fragmenting such previously large markets as the United States and raising the costs of producing for sale across it. A new generation of multilateral environmental agreements has established an additional set of regulations at the international level, and often incorporated trade restrictive measures to enforce compliance with them.

At all levels, from the local to the international, regulation has come to focus not only on the characteristic of products entering a market but on the methods by which they are produced, processed, distributed and disposed of anywhere in the world and all along the value chain. The traditional focus on product characteristics has thus been joined, in a cradle-to-grave vision, by regulatory action over the entire product cycle, from production and processing methods, to waste reduction and product disposability. Firm reputations and liabilities can now be strongly affected by the practices of their suppliers and partners. The emergence of a total systems

approach to environmental protectionism has intensified the clash among once separated industries, such as autos and oil, about who could and should best bear the additional environmentally-enhancing costs. This maze of multilevel, rapidly changing, ever more stringent, and far-reaching regulation imposes major costs on those firms unable to respond before their competitors to the harsh demands of this new environmental regulatory world.

Accompanying these developments in the arena of environmental regulation are similarly far-reaching changes in the world of international business. The advent of internationally integrated production and just-in-time inventory processes means firms need to be able to import and export freely and without interruption across international boundaries. This requirement compounds the costs, which even minor regulatory detours or delays can impose, and simultaneously generates new incentives for convergent or at least compatible environmental regulations across jurisdictions. Moreover, the expansion in multinational enterprise and international business alliances has further increased the cost of national regulatory protectionism and reinforced these pressures for compatible production standards (Dunning, 1993; Rugman, 1996). Finally, the advent of competition on a global scale, often from firms of much larger size, has underscored the need for a large protected home (or now multi-country regional) market to amortise fixed costs and build the required minimum scale.

Governing this clash between the world of more complex environmental regulations and global competitive pressures over the past decade is a third new development – the advent of potent international institutions with trade and environment responsibilities. When, as with NAFTA, such institutions are accompanied by strong powers, institutions and organisations that embrace both trade and the environment, they can provide a common forum and reference point for trade-environment communities to interact and their conflicts to be settled, managed, or event protected. And if, as in NAFTA, they provide for the direct participation of firms and environmental groups, as well as governments, from all participating countries, they can foster the development of complex transnational coalitions, and an emergent sense of common interests *vis-à-vis* the outside world.

The more heavily and innovatively regulated, internationalised and institutionalised world of the 1990s presents firms not only with complex new obstacles. It also offers them new opportunities, in both their corporate and political strategies, to circumvent potential barrier and increase their competitiveness on a larger scale. Of interest is the vastly expanded array of corporate responses that firms, following a strict business logic, can now employ.

Traditional corporate strategies

Traditionally, firms facing the classic challenge of environmental regulatory protectionism in their major export market abroad have had a limited array of often unsuccessful strategies to employ. The first, often instinctive response was to secure from the foreign regulatory authority a certificate of equivalency or another form of *ad hoc* exemption for a specific shipment or product. Such a response of

ad hoc exemptionalism was tried unsuccessfully by Lactel, at an early stage of the UHT case (Rugman *et al*., 1997; Vogel and Rugman, 1997).

A second response, attractive when the foreign regulations were prohibitive for market entry, was to withdraw to the domestic market and seek replacement customers at home. Such a strategy of domestic withdrawal was an appealing second best option for smaller firms with a large domestic market, and for firms with products at early stages of the product cycle where domestic demand could be expected to rapidly expand.

A third response, attractive to firms which wished to preserve their export market, was to pay the incremental cost of the foreign regulation, hoping that by expanding production through foreign sales they could lower their unit costs and restore net profitability in their foreign sale. Such a response was attractive where the cost of the foreign regulation was modest rather than prohibitive, potentially transitory (while undergoing further scientific testing, for example), and where the export market was rapidly growing and serviced by few competing firms at home or abroad. In sunrise sectors, where the regulation stemmed from early government regulatory authority backed by few national producing industries and allied NGOs, such a strategy could prove to be a successful one over the longer run. This strategy was followed on the part of the US beer producers in the Ontario beer case (see Rugman and Soloway, 1998).

A fourth strategy was to alter one's product or even production standard to meet the foreign regulation of the moment. This was an attractive option where the costs of the adaptation was low (perhaps restricted to a single low value input), where firms were installing new capacity in a product line dedicated to the export market and segmented from the rest of the firms production capacity, and when the foreign market offered high profitability, a high percentage of the firms sales, and confidence that the existing regulation would last for a long time (over the course of which incremental costs of adjustment could be amortised). Such a strategy of product/production alteration was evident in the case of New Brunswick blueberries, where the Canadian producer switched from the use of the pesticide dimenethoate to that of higher cost but still affordable imidian, in an effort to meet the US regulation (Rugman and Soloway, 1998).

A fifth strategy was to shift to the production of alternative, even closely allied products that did not face the specific regulatory restriction. This was particularly attractive when the abandoned product, relative to the alternative, was a mature, low-value-added product whose market share might be in a state of long term decline. It was also attractive when the foreign regulatory barrier was highly product specific, and difficult to change or expand. This strategy of product alternatives was employed, successfully for a few years, by Canadian producers in the softwood lumber case.

A sixth traditional strategy was market diversification – shifting sales to an alternative export market where the regulatory barrier did not exist. This was particularly attractive to mature export capable firms able easily to overcome barriers of distance and language, to seek new customers in less or differently regulated markets. The diversion of newsprint exports from California to Asia was a strategy

considered on the part of the British Columbia forestry firms when they were denied access from the California market (Vogel and Rugman, 1997).

These traditional strategies were essentially reactive rather than proactive – they responded to the introduction by foreign governments of new or different environmental regulations. Although they depended on a calculation of how those and related regulations were likely to change in the foreign government, in other export markets and at home, the dominant assumption was that a foreign government would act without catalysing a sequence of strategic moves by other players. This would include other governments with regulatory powers, and firms and NGOs wishing to adopt or circumvent successful techniques. These traditional strategies are thus poorly suited to the new world of rapidly changing environmental regulation and strategic regulatory protectionism, extensive internationalisation of business, and strong international institutions to constrain national regulations and create international regulations of their own.

New corporate strategies

The first of these new strategies, developed in the work of Porter (1990), is to readily accept and produce at home to the highest environmental standards in one's domestic market, in order to have the unique first mover advantage in export markets which are calculated to be moving to ever higher levels of environmental regulation (Porter and van der Linde, 1995). Such a strategy, however, is available only to relatively large firms who can afford the initial additional costs, who have long time horizons, who can use their green production as an advantage at home among consumers to capture the domestic market and secure the required scale, and who have a large unfragmented national market to allow the minimum required scale (Rugman 1995). They are also appealing when the level rather than the form of regulation is at issue, when there is a single future standard that is probable and widely accepted, when regulations rise by a unilinear rather than leapfrogging process, and when technology forcing innovation is likely.

The second strategy, see Vogel (1995), is to meet the high environmental regulations prevailing in the largest export market. This is the strategy that takes advantage of the 'California effect', under which German automotive producers at home and abroad met the higher environmental standards in their large California market, confident that the environmentally pioneering California regulations would eventually be adopted throughout the United States and in key markets around the world. Such a strategy offers strong first mover advantages in a world where markets, and higher environmental regulations, are becoming global. It involves significant risks, however, when several subfederal jurisdictions and countries are seeking to become the global environmental pioneer and when their differing regulatory approaches threaten to fragment the prospective global and existing national marketplace.

A third strategy is to move production facilities closer geographically to an export market, in order to minimise the transportation costs, and to better absorb the temporary border delays that environmental regulations sometimes cause. This

strategy is attractive to home-based exporters from countries such as Canada and Mexico *vis-à-vis* the United States, where a firm's home country is contiguous to the major export market and to the major consumers within it. Its logic is seen in the many operations that have opened and expanded, both before and after NAFTA, in the *maquiladoras* along Mexico's northern border. It was a potential strategy available to Canadian newsprint producers faced with the task of transporting recycled newsprint from major urban centres in California and elsewhere in the United States back for recycling in Canada (Rugman, 1995). It is also attractive where just-in-time inventory methods are in use, and where there are additional advantages, for learning and marketing purposes, in being geographically proximate to customers.

A fourth strategy is to transform oneself from a home-based exporter into a domestic producer, by abandoning the newly closed export market and selling one's product as an input to another domestic producer whose products are free to enter the foreign market. This is the strategy pursued by Canada's leading steel producer, Stelco, when its steel exports to the United States were hit with a succession of antidumping duties. It diverted its product to supplying Canadian auto parts manufactures such as Magna who were free to ship to the United States. Here US protectionist action had the unintended effect of inducing a successful 'further processing' industrial strategy in Canada (Rugman and Anderson, 1997).

A fifth strategy is to transform oneself from a home-based exporter into a home-base multinational enterprise by opening production facilities in the country whose market is now closed to exports by the environmental regulation. NAFTA's investment guarantees allow firms to adopt this strategy with much greater assurance than before. This strategy is attractive in cases, such as the recycling cases, where geographic proximity is an advantage, and where the costs of a greenfield or take-over investment, including those of regulatory compliance, are not onerous. The process fostered by the use of this strategy is the opposite of the widespread fear of regulatory refugees closing facilities in high cost environmental sanctuaries for relocation in pollution havens where environmental regulations and thus business costs are lower. Under this strategy firms move, but to jurisdictions with higher environmental regulations. It is likely to be most evident where countries employ environmental regulations that are easy and inexpensive for domestic firms, but difficult and costly for foreign firms, to meet.

A sixth strategy is when firms from both the importing and exporting country make an agreement regarding minimum prices in exchange for the withdrawal of anti-dumping actions or harassment through discriminatory environmental regulation. Such agreements exist outside the NAFTA institutional process. This has been especially prevalent in the agriculture sector, where in the case of Mexican tomatoes exports to Florida, such an agreement was made. The Florida tomato producers agreed to withdraw their anti-dumping actions and cease harassment over environmental and packaging standards in exchange for voluntary export quotas and minimum price floors on the part of the Mexican tomato producers.

Traditional political instruments

Traditionally, the first response of firms faced with regulatory protectionism has been litigation. This requires mobilising the power of one's home government to take up one's case and be an advocate in the national trade law system of the country imposing the regulation. For example, in the UHT case, Lactel mobilised the home, Canadian government at the official and diplomatic level to help it secure a certificate of equivalency for its product (Rugman and Soloway, 1998). This is essentially a passive strategy, in which the firm and its government are responding to the procedures and schedule of the foreign trade law system. As the experience of Lactel in the UHT milk case illustrates, the strategy of litigation, even when successful, can take so long, and involve such expense, that serious competitive damage is done. More costly still is the fear, borne by repeated experiences of such 'process protectionism' against the firms' products, that future exports will be blocked to the point that the firm is deterred from pursuing future exports (Vogel and Rugman, 1997).

A second political strategy is lobbying, with the aim of inducing the foreign government to allow such discretion as its national trade law system exists to provide at least an exemption in such a case. Such a claim can often best be made by one's home government, as an exercise in 'exemptionalism' (Kawasaki, 1998; Cuff and Granatstein, 1972). Firms can reinforce their capacity to pursue such a strategy by retaining local legal counsel with political access in the foreign country, or by maintaining permanent corporate representation in the foreign capital (Gotlieb, 1984). Here home-based MNEs have an advantage over home-based exporters in that they can employ their host country managers to gather intelligence about, and intervene as nationals in the legal and related political processes of the regulating government.

A third, more active strategy is coalition-building – constructing and activating a broad coalition of interest groups within the regulating country and in third countries to secure a modification or suspension of the objectionable regulation. For example, in the softwood lumber case, and earlier related cases such as shakes and shingles (Rugman and Anderson, 1987), Canadian firms mobilised the US consumers' interests. This included the US homebuilders associations (to whom the product was sold), those financing the US housing purchases, and consumers wishing to purchase homes. All of these groups had a vested interest in the low input prices and easy availability of Canadian lumber. Reciprocally, in the Ontario Beer case, Alcan (as a Canadian supplier to US firms exporting their beer in aluminium cans into the Ontario market), was mobilised against the Ontario environmental regulation (Vogel and Rugman, 1997). Transnational MNEs, with a presence in many countries, are often best positioned to employ this strategy as they can more easily mobilise third-party governments to join an intervention against the regulating government.

A fourth strategy is high-level diplomacy. This involves mobilising one's home government, at the leader, ministerial or senior official level, to pursue the case on an intergovernmental basis, and perhaps link it to other issues in an overall bilateral

relationship. Such a strategy is reinforced by the finding that in the Canada–US relationship, Canada tends to prevail in issues dealt with at the summit level and linked to the overall state of the relationship (Nye, 1974). Such a strategy can be pursued as a matter of segmented problem solving in the special relationship, an approach pursued with the joint Canada–US study by agricultural officials in 1995 in the UHT case. Or it can be done as a matter of pure intergovernmental bargaining. For example, in the softwood lumber case, following its failed arbitration through the FTA, including two Extraordinary Challenge Committees, the two national governments negotiated outside the trade law system to arrive at the 1996 Softwood Lumber Agreement with its own dispute settlement procedures (Rugman and Anderson, 1997).

A fifth strategy, attractive when the foreign government will not easily adjust, is subsidisation – seeking a subsidy from one's home government to meet the cost of the foreign environmental regulation. It does raise the danger that such 'generally available' subsidies will trigger further countervailing duty action by the regulating governments. This is especially true in the case of the United States, where such duties are imposed on the gross amount of the foreign subsidy rather than the net difference between foreign subsidies and US subsidies to its home firms (Rugman and Anderson, 1987, 1997). This strategy is attractive when the size of the required subsidies is not large, when one's home government has an available fiscal surplus, and where there is a legitimate case for infant industry, automatically sunsetted subsidisation (as with a one-time grant for environmental facilities for new plant construction).

A sixth strategy is retaliation – having one's home government impose mirror image regulatory restrictions of equal or greater magnitude on the imports from the country practising the initial regulatory protectionism. This use of this strategy is evident in the counter-retaliations over potato inspections along the Maine–New Brunswick border and in countless other similar cases between the United States and Canada (Rugman and Soloway, 1998). It is likely to be efficacious when the counter-retaliation is carefully targeted against politically consequential constituencies in the other country, or where the market share of the imports and thus leverage of the retaliating country is larger than that that of its rival.

A seventh strategy is one of convergent national adjustment, that is, where firms intervene with their home government to adjust its national regulations unilaterally to correspond with those of a major foreign government to which the firms exports. This is essentially a strategy of acquiescence, based on a calculation that a single international regulatory regime regardless of content is more beneficial through the reduction of transaction costs than nationally diverse systems of whatever character. This strategy stands in contrast to Porter (1990) where a firm seeks high home government regulation in advance of those foreign markets, and in contrast to Vogel (1995) where a firm seeks to meet foreign government standards in advance of its own home government. Convergent national adjustment was the strategy pursued by Canadian auto manufacturers in demanding that the Canadian government change its regulations with respect to MMT in gasoline to conform to the United States standard which did not allow MMT (Rugman and Soloway,

1998). This strategy rationally appeals to MNCs for whom international transaction costs can constitute a primarily regulatory barrier.

Whatever strategy, or blend and sequence of such traditional strategies is employed, all are heavily dependent for their use and success on the willingness of one's home government to take up the case of a single firm, to pursue it effectively with a foreign government and within a foreign country. Such dependence on one's home government can be a problem. At home, a firm might be too small and politically inconsequential to secure home government support, by virtue of its size, number of employees, regional location, campaign contribution record, or image and status as not being a national champion. Firms may not be able to mobilise sector-wide coalitions of other firms at home if the latter perceive that they can secure a competitive advantage from avoiding collective action (if they are less dependent upon exports to the regulating country's market). Domestic public opinion may make it difficult for a national government to vigorously pursue a national firm's cause, as indicated by the case of aboriginal leghold traps or Canada's east coast seal hunt (the conduct of which offended domestic animal rights and environmental groups).

Looking abroad, a government may have limited leverage with a much larger foreign government, particularly at times of cool overall relations. And the interests of a particular firm might be traded off for higher political issues or to maintain good overall relations with a foreign government.

New political instruments

The new political instruments and strategies available under complex institutional responsiveness circumvent many of these obstacles. They are far less dependent for their use and success on the will, skill, and size of a firm's home national government. They still involve political action, but focus more directly on using other political forums and actors, particularly those opened by the new array of international institutions. Whereas the old strategies (rooted in mobilisation by national governments) contain a bias toward entering into and escalating intergovernmental conflict, the new strategies (based in international institutions), begin with the rules-based third party treatment of firm-government conflict and contain a built in bias towards ever stronger forms of transborder co-operative action.

The first of these new political strategies is international dispute settlement. This involves taking one's dispute directly to the new NAFTA network of dispute settlement mechanisms. Like their FTA predecessors they offer a rules-governed international forum with specific mechanisms for anti-dumping and countervailing duties cases (Chapter 19), and for general disputes (Chapter 20). Yet unlike the FTA they also offer, under Chapter 11, two mechanisms for investment disputes, and under the NAAEC, three mechanisms (Chapter 13, 14–15 and Part 5), for dealing with environmental disputes. Most innovatively, two of these six mechanisms – NAFTA's Chapter 11 and NAAEC's Articles 14–15 – allow firms direct access to initiate and pursue cases, without involving their home government. The NAAEC's Article 13 enables firms to lobby the CEC Secretariat directly to initiate

an investigation. The early widespread use of Chapter 11 for environmental cases, and that environmental cases constitute virtually all of the Chapter 11 cases to date, plus the absence of environmentally related cases among those dealt with by the more FTA-like Chapter 19 and 20 mechanisms shows the importance of this right of direct firm access. Further, the widespread use of NAAEC's Article 13 and 14–15 cases, and the lack of any action under Part 5 (which requires government to government action) underscores the point. Firms, and their competitors are no longer hostages to home governments. They can now move to redress national and local environmental regulatory protectionism directly at the international level.

The second of the new political strategies is dispute management and prevention. The mandate and the work of the NAFTA institutions and the CEC provide an incentive and capacity to deal with disputes before they automatically proceed to expensive and visible litigation through dispute settlement. Through the role of the ministerial councils and their subordinate official-level institutions and stand-alone secretariats, disputes can be politically managed in ways that prevent them from escalating into major intergovernmental conflicts, or from arousing domestic political pressures that would further impede transborder trade. These institutions also have the power to act to prevent such disputes from arising in the first place. This capacity has proven its value already, specifically in regard to major regulatory protectionist issues relating to safety, notably in the US–Mexican trucking area. The role of the NAFTA co-ordinators and, since 1998, the NAFTA deputy ministers forum, has substantially reinforced this capacity for dispute management and prevention.

A third strategy, which extends the impetus for dispute prevention from conflict avoidance into the realm of actual co-operation, is the use of the NAFTA institutions for communication and capacity-building. Through NAFTA institutions, firms can learn about the regulations, certification procedures and accreditation, and regulatory development systems in other countries. Firms also have a NAFTA guaranteed right to comment on draft regulations and thus shape their content. NAFTA's institutions encourage firms to build regulatory capacity in partner countries, thereby ensuing that foreign regulations are similar to home country ones in the consistency and predictability of their application. Such capacity building can also aid foreign suppliers and the general business infrastructure. In both cases, the development of personal networks with those in partner's national regulatory systems can do much to alleviate misunderstanding, build trust and lower the costs of commerce.

A fourth strategy is regulatory convergence – using NAFTA institutions to have national and local regulations move over time to become more compatible or similar. There are many ways this harmonisation process can take place, from mutual recognition of national standards, through minimum standards and procedures to the negotiation and acceptance of common standards (Esty and Geradin, 1997). Such a process is easier to undertake and complete in areas of new regulatory activity, where no partner country has existing regulations and the interests that lie behind. It can vary in the level of the common standards aimed at (high versus low), the dynamics of movement (negotiated versus incremental; hegemonic

versus mutual adjustment), the form (labelling, product, process or disposal standards), the scope (regional versus multilateral) and the speed (including differential phase-ins).

A fifth strategy, facilitated by NAFTA institutions, is to form transnational coalitions to secure the convergent regulations one favours. Because the NAFTA institutions bring together firms, ENGOs and other stakeholders from all participating countries in their work, they ease the task of assembling transnational coalitions to overcome national obstacles backed by weaker, national firm coalitions. Broader coalitions can also widen the range of benefits available. Over time, this process can lead firms and their governments to redefine their interests, and to formulate business strategies in fully regional terms.

A sixth strategy is to engage in multilaterally-oriented coalition-building. This involves constructing coalitions across all communities and countries within the region to create a stronger North American constituency, aimed at securing the broader multilateral standards that favour North American industry interests. This strategy has been evident in the activities of firms in the chemicals industry.

A seventh and final strategy is to engage in regional harmonisation, with or without a multilateral orientation, outside of the NAFTA institutions or even non-NAFTA intergovernmental forums. Such international, voluntary, private sector, standardisation can be seen in the activities of the automotive industry, for whom the work of the relevant NAFTA institutions is distinctly secondary. It can also be seen, outside of the environmental domain, in the activity of the North American steel industry, where, in 1998, 48 steel firms have formed a region-wide association and have begun to act against the threat of low cost steel flowing into North America from currency-weak Asia and Russia.

The impact of the NAFTA institutions on firm behaviour

An expanded array of corporate and political strategies available under conditions of NAFTA's complex institutional responsiveness has been documented. Which are those likely to be, under rational calculation, most appealing to different types of firms? The advent of free trade agreements such as NAFTA, with strong trade and investment liberalisation provisions, plus a thick network of institutions for shaping regulations and managing disputes, offers firms a new arsenal of opportunities and instruments to maintain and enhance competitiveness. These environmentally-related institutions have thus far operated effectively to facilitate the access of Canadian and Mexican firms to the US market, and that of US firms to their partner's markets. They have also served to augment the ability of North American firms to penetrate global markets, and to enhance the North American environment at the same time. The basic rationale for particular instruments to be employed by these different types of firms is as follows.

In the first instance, domestic firms can take advantage of the new NAFTA institutional network in several ways. Most directly, small domestic firms can benefit from the direct contracting practices of the CEC itself. The CEC has allocated its modest budget with strict attention to equal disbursements among the three mem-

ber government contributors, often through the formula of contracting enterprises from the three countries to work together on a particular project. Such certification by an international organisation and development of a network of affiliates in the other two NAFTA countries, even at a very modest level, increases the export readiness of small domestic firms, helps to transform them into home-based exporters, and gives them an international network of firms with which they might supply or form co-operative alliances in the future. For small, domestic firms such a lowering of transaction costs and provision of a *de facto* guarantee of international governmental approval can be of considerable value.

The work of the NAFTA institutions more generally, particularly through its involvement of firms, has a substantial effect in bringing larger domestic firms and home-based exporters into contact, not merely with foreign distributors and customers, but with similar firms or potential competitors. From such concentrated contact can flow a greater sense of export opportunities, familiarity with local regulations, and the identification of local lobbying affiliates. It can also breed the trust and information required to forge larger strategic alliances.

NAFTA can assist home-based exporters in another way. Such trade agreements, by opening vastly larger markets with different, and in the case of Mexico at times less stringent, regulations give formerly domestic producers and their home country much larger markets for their existing products, and thus the resources and skills to meet the intensifying regulatory demands from abroad. At a minimum, NAFTA's trade liberalisation provisions mean firms could find new markets abroad (Mexico) where their products meet the regulatory requirements, and thus insulate themselves from and reinforce their ability to compete with regulatory protectionism in a single large market such as the United States. NAFTA thus offers home-based exporters fully focused on the United States the strategy of export diversification.

NAFTA's trade liberalisation provisions, and its innovative guarantees for foreign direct investment make it easier for firms to move production abroad, to service the former export markets, and the former home market. NAFTA thus gives home-based exporters the option of becoming home based MNEs, at least on a regional scale. They can move existing and new production to lower cost jurisdictions, or they can move into the US market itself, and thereby better learn about and lobby for the national and subnational regulations they can meet. In doing so, they place, as NAFTA's critics correctly identify, a market-based check on excessive enthusiasm for national environmental regulatory protectionism, especially the form that could harm firms in the global marketplace.

In addition, the NAFTA Agreements' environmental provisions and institutional mechanisms give firms a much enhanced array of possibilities. High and costly environmental regulations at home can force firms to look for investment locations abroad, where other production costs could be higher, distance from high value markets greater, and access to those markets more uncertain. In such a situation, of particular interest to experienced home-based exporters with several foreign markets and to home-based MNEs are those CEC programs directly focused on corporate contributions to environmental enhancement, notably the

development and spread of voluntary environmental management systems such as ISO 14000, and support for pooling environmental technologies of proven value in Mexico to enhance their export into the Americas as a whole. Their participation with ENGOs can also lead to shared interests either in sustainable development opportunities, or in green protectionism at a regional rather than the national or subnational level.

By far the most important impact of the NAFTA institutions, especially for the home-based MNEs that dominate North American trade and investment, is their work in: 1) constraining the emergence and use of green protectionist regulations by national, subfederal and local governments within the NAFTA community; 2) facilitating regulatory convergence across the region; and 3) fostering a single North American regional voice to combat such protectionism in external regions such as the EU and build a broader, more open multilateral regime.

A substantial number of the environmentally-related trade disputes in the region (affecting domestic firms, home-based exporters and home-based MNEs), especially prior to NAFTA, arose from state and provincial regulations (Vogel and Rugman, 1997). While such cases have thus far concentrated on agricultural and natural resource industries, pressures for state-level automotive emission standards (beyond California) and automotive inspection and maintenance programs threaten to bring such subfederal regulatory escalation and proliferation to the manufacturing industries at the heart of the NAFTA economy.

For home-based exporters and MNEs it is often the uniformity, stability and form rather than the level of regulation which is the central interest. Although MNEs are conventionally thought to be in favour of regulatory diversity across countries, so that they can exploit their comparative advantage by relocating to produce for export to global markets from the locations where their costs are the lowest, in practice it is costly to shift production to such locations. Moreover the high degree of intra-corporate trade and management integration provides further incentives for firms to favour stable or slowly changing uniformity over rapidly changing diversity.

The NAFTA regime and institutions, in practice, have not led to a downward harmonisation that have forced subfederal entities to reduce their environmental standards, where they are higher than the national or international norm (Kirton, 1998; Orbuch and Singer, 1995). Although allegations of a deterrent 'chill effect' persist (by which the NAFTA disciplines and institutions prevent subfederal entities from further increases and thus differentiation in their environmental regulations), the record of dispute settlement thus far suggests the incentives are otherwise. For while the CEC has acted against subfederal jurisdictions (such as British Columbia) for the alleged non-enforcement of existing environmental regulations, the NAFTA dispute settlement mechanisms have not yet moved against those seeking to set environmental regulations higher (apart from the MMT case in process against the Canadian federal government). The NAFTA institutions and the CEC thus assist home-based exporters and MNEs with their most costly market access problem of subfederal regulatory proliferation, with little cost to existing levels of regulatory protection and some benefit against those states and provinces tempted to relax environmental enforcement.

It is here that the NAFTA institutions, implementing the NAFTA rules on standards, have been most effective. By bringing NAFTA disciplines to many areas of state and provincial regulatory activity, while restricting the direct participation of state and provincial officials in them, the NAFTA institutions in their composition and practice create a bias in favour of standards that are trade friendly, national and regional in application (Orbuch and Singer, 1995).

More generally, the NAFTA rules and institutions have assisted home-based exporters and MNEs from NAFTA's two smaller countries with the major threat they face, namely loss of market access to the United States. With unilateral environmental regulatory action by the national government of the United States, entrenched in a Vogel (1995) type 'baptist–bootlegger' coalition, there is the threat of denial of access to the large and vital US market to firms who must export their products there to survive. The most analytically clear instances arise where the US government has embedded in its national laws trade sanctions against foreign firms allegedly not abiding by the practices prescribed in US regulations.

These cases follow a standard pattern, in which, first, US environmentalists seek and secure a higher US environmental standard (or support an international one). Second, US industry resists its imposition on US industry (or demands its imposition on external competitors from foreign countries). Thirdly, a 'green and greedy' coalition is formed to secure unilateral trade sanctions as an enforcement mechanism. Fourthly, the industry proceeds with its environmental partners to seek the internationalisation of this standard (DeSombre, 1995). Such a process, which might be termed the 'Washington effect', explained earlier in the chapter, as distinct from the 'California effect' (Vogel, 1995), opens up two major strategies for foreign firms. The first is proactively forging alliances with US firms in the first stage to prevent such national environmental regulations from coming into force (or doing so in such a way that exempts Canadian firms from its application) and thus provides privileged access into the US market.

The second is to pre-emptively create an international standard (at or near the higher US level or acceptable to the dominant US industry–environmental–government alliance) before the damaging threat of unilateral sanctions emerges. The presence of a three country regional forum, in which the US government feels comfortable, and with effective institutions to devise, implement and enforce such standards, is a major asset, particularly to home-based exporters in Canada and Mexico. It helps to speed up the second strategy, given the slowness of the broad multilateral process and the latter's anti-environmental bias arising from the large number of developing countries exercising a drag effect.

At the firm level, the high degree of intra-firm integration between Canada and the United States provides MNEs with a strong incentive to eliminate potential trade sanctions that might impede their production systems. At the same time, the presence of a sanctionist regime (through fines, trade penalties or domestic court action), has an impact in inhibiting US firms from lowering the effective environmental regulatory burden in their prized domestic market and reaping the rewards which may ensue. It is thus a victory for Canadian and Mexican domestic firms subject to US import competition, and home based exporters (competing

against domestic US rivals in the US market) at the expense of their US equivalents, while having an essentially neutral effect on US, Canadian and Mexican home-based MNEs.

Looking ahead, the greatest value of NAFTA's environmentally-related institutions will come in altering the interests of their governments and firms, and enhancing their capacity to advance common North American interests in broader multilateral forums. The trend toward broadly multilateral standardisation will primarily benefit US MNEs whose broadly dispersed international markets and production create an overwhelming interest in a single, high level environmental standard worldwide. The trend toward a regional caucus and common North American standards, distinct from that of the rest of the world, will primarily benefit Canadian and Mexican MNEs and exporters, as their major markets and production sites are primarily located in the United States and North America itself.

Conclusions

As firms operate in a more complex environmental regulatory system, they are faced with numerous challenges and opportunities. Firms at all levels need to take account of the new environment for complex institutional responsiveness, that is, careful strategic planning is required in order to navigate around these challenges and make the most of these competitive opportunities. Traditional corporate and political strategies available prior to the advent of the major multilateral and regional trade liberalisation initiatives are no longer the most effective course of action. They tend to be reactive rather than proactive, and fail to take account of the realities in which firms must operate today where business is internationalised, environmental regulation is rapidly changing and a new international institutional structure exists.

Conditions of complex institutional responsiveness offer a much broader array of both corporate and political strategies, which take account of the new regulatory, competitive and institutional processes at work. In the case of NAFTA, its new set of international institutions charged with trade and environment responsibilities has provided multiple opportunities for firms which they are beginning to utilise. The NAFTA institutions have had particular success in constraining the emergence of 'green' protection at all levels of regulation, facilitating regulatory convergence, and developing a North American 'voice' to combat protection beyond its borders and influence the development of standards in international bodies. In the future, sectors where there is a high degree of both region-wide integration and institutional-based regulatory convergence will be best poised to take advantage of the new environment for complex institutional responsiveness.

The array of corporate and political strategies discussed in this chapter can be generalised beyond the North American experience. The shifts from traditional to new corporate and political strategies are relevant for firms operating under the GATT/WTO, the EU, FTAA, and to some extent APEC. Where the conditions for complex institutional responsiveness are present (as they are embodied in NAFTA), there is a rational incentive on the part of firms to use the new, rather than old, strategies of complex institutional responsiveness. Complex institutional respon-

siveness is not the fate of North American and European firms only. Rather, it is expected that the conditions of complex institutional responsiveness will develop and gain momentum over time. Thus one can expect that firms will embrace these strategies as new institutions develop, as environmental concern becomes more pressing and the conduct of business becomes ever more international.

References

Cuff, R.D. and J.L. Granatstein (1972) 'Canada and the perils of exemptionalism', *Queen's Quarterly*, 79: 673–81.

DeSombre, E. (1995) 'Baptists and bootleggers for the environment: the origins of United States unilateral sanctions', *Journal of Environment and Development*, 4: 53–75.

Dunning, J.H. (1993) *Multinational Enterprises and the Global Economy*, New York: Addison-Wesley.

Esty, D. and D. Geradin (1997) 'Market access, competitiveness, and harmonisation: environmental protection in regional trade agreements', *The Harvard Environmental Law Review*, 21 (2): 265–336.

Gotlieb, A. (1984) 'Canadian business representation in the U.S.', *Behind the Headlines*, 42.

Kawasaki, T. (1998) 'Managing macroeconomic relations with the United States: Japanese and Canadian experiences', in M. Fry, J. Kirton and M. Kurosawa (eds) *The North Pacific Triangle: The United States, Japan and Canada at Century's End*, Toronto: University of Toronto Press.

Kirton, J. (1998) 'NAFTA's foreign direct investment and economic integration: the case of Canada', paper prepared for an OECD Seminar on Migration, Free Trade and Regional Integration in North America, Mexico City, January 15–16.

Nye, J.S. Jr. (1974) 'Transnational relations and interstate conflicts', *International Organisation*, 28(4): 961–98.

Orbuch, P. and T. Singer (1995) 'International trade, the environment and the States: an evolving state-federal relationship', *Journal of Environment and Development*, 4: 121–44.

Porter, M.E. (1990) *The Competitive Advantage of Nations*, New York: Free Press/Macmillan.

Porter, M.E. and C. van der Linde (1995) 'Toward a new conception of the environment-competitiveness relationship', *Journal of Economic Perspectives*, 9 (4): 97–118.

Rugman, A.M. (1995) 'Environmental regulations and international competitiveness: strategies for Canada's west coast forest products industry', *The International Executive*, 37 (5): 451–65.

Rugman, A.M. (1996) *Multinational Enterprises and Trade Policy*, Cheltenham: Edward Elgar.

Rugman, A.M. and A.D.M. Anderson (1987) *Administered Protection in America*, London: Routledge.

Rugman, A.M. and A.D.M. Anderson (1997) 'NAFTA and the dispute settlement mechanisms: a transaction costs approach', *The World Economy*, 20 (7): 935–50.

Rugman, A.M. and J.A. Soloway (1998) 'Corporate strategy and NAFTA when environmental regulations are barriers to trade', *Journal of Transnational Management Development*, 3: 231–51.

Rugman, A.M., J. Kirton and J.A. Soloway (1997) 'NAFTA environmental regulations and Canadian competitiveness', *Journal of World Trade*, 31 (4): 129–44.

Vogel, D. (1995) *Trading Up: Consumer and Environmental Regulations in a Global Economy*, Cambridge, MA: Harvard University Press.

Vogel, D. and A.M. Rugman (1997) 'Environmentally related trade disputes between the United States and Canada', *The American Review of Canadian Studies*, 27 (2): 271–92.

11 Green and producer lobbies

Enemies or allies?

Paola Conconi

The purpose of the analysis in this chapter is to understand how the presence of green and producer lobbies can affect the political determination of trade and environmental policies.

Recent events in the United States have illustrated the extent to which citizen groups condition trade and environmental policies, both at the national and multilateral level. On the trade side, the creation of the North American Free Trade Agreement (NAFTA) initially encountered the resistance of business, labour and environmental groups (VanGrasstek, 1992). By pledging in an environmental side agreement,[1] the White House was able to win the support of at least some environmental groups and obtain the fast track authority to negotiate the trade agreement without a line-by-line veto from Congress.[2] More recently, environmental groups have joined forces with protectionist industries and labour groups to launch a fierce campaign against further trade liberalisation, which has caused the breakdown of the new round of WTO negotiations in Seattle.[3] Industry and green lobbies have been extremely influential also on the environmental side. On some issues, such as multilateral emissions cuts, they have held different positions. For example, the strength of the producer lobby has caused the USA to abandon the Kyoto Protocol, after the assumption of office by President Bush. On others, such as the compliance of foreign legislation with American environmental standards, their objectives have often coincided.

This chapter attempts to shed some light on the relationship between green and producer lobbies. In particular, we wish to address the following questions: when will their interests over trade and environmental polices be aligned and when will they diverge? What will be the unilateral and cooperative policies selected by politically minded governments? When will policy coordination be efficiency enhancing?

Understanding the nature of the relationship between lobby groups is important for two reasons. On the positive side, it can help us to explain observed trade and environmental policies. On the normative side, it can provide some guidance on how to construct efficient policy mechanisms in the presence of political distortions.

In Conconi (2002), we studied how green lobbying can influence the determination of trade and environmental policies when countries are large and emissions are transboundary. Here we extend the analysis to a situation in which both producer and environmental interests are organised.

To examine the relationship between interest groups and policy-makers, we adopt the common agency model pioneered by Bernheim and Winston (1986) and applied to trade policy by Grossman and Helpman (1994, 1995). The common agency problem arises when many principals deal with the same agent. A national or supra-national government is the agent who sets trade and environmental policies. Green and producer lobbies act as principals and confront the government with contribution schedules, namely functions describing their political contributions contingent on the chosen economic policies. These can be interpreted as legal campaign contributions, support demonstrations, or simply bribes. The timing is that first lobbies simultaneously commit to contribution schedules, and then the government, having observed these schedules, sets trade and environmental policies. The implicit objective of incumbent politicians is to be re-elected. They trade off the political support that comes from heeding interest groups' demands against the alienation of voters that may result from the implementation of socially costly policies.

A key feature of our model is that the countries considered are large in that they are able to affect world prices. This implies that a unilateral increase in domestic pollution taxes or decrease in import tariffs generate *emission leakages*, they shift the terms of trade away from the implementing country, leading to an increase in emissions by its trading partners. If emissions cross borders, the increase in foreign pollution will have a negative environmental impact from the point of view of the domestic residents.

We characterise the policy outcomes and the relationship between lobbies in three alternative policy regimes: one where governments control both trade and environmental policies; one in which they are restrained to the use of environmental policy by an existing free trade agreement; and one in which trade policy is the only available instrument. We find that, in the presence of emission leakages and transboundary spillovers, the relationship between green and producer interests over trade and environmental policy is ambiguous. If instead pollution is local and/or the emission leakages are eliminated through the combined use of trade and environmental policy instruments or through international policy cooperation, green and producer lobbies will unambiguously be either enemies or allies.

This chapter also contributes to a growing literature which examines the influence of interest groups on policy-making.[4] Most existing studies, however, focus on a single policy instrument.[5] To the best of our knowledge, ours is the only study looking at the role of green lobbies on the *joint determination* of trade and environmental policies in large open economies.

The issue of the link between the trade policy regime and stringency of environmental regulations has been recognised in number of papers. A study by Perroni and Wigle (1994) shows that, given the level of environmental regulations, trade policy has little impact on the quality of the environment. Husted and Logsdon

(1997) find instead that the NAFTA agreement has led Mexico to strengthen its environmental policies. On the theoretical side, Fredriksson (1999) examines a scenario in which environmental and industry interest lobby groups influence the determination of pollution taxes in sectors protected by tariffs. The level of protectionism is exogenously determined. The main result of his analysis is that the level of political conflict on environmental policy falls with trade liberalisation. Schleich (1999) studies the joint determination of trade and environmental policies. The government is assumed to have a single or a variety of domestic and trade policy instruments to address production or consumption externalities and to obtain political contributions from producer lobby groups. He shows that, in the presence of both trade and environmental distortions, inefficient trade policies can lead to higher environmental quality than more efficient domestic policies. Differently from our analysis, both Fredriksson (1999) and Schleich (1999) focus on a small economy and on local environmental problems, thus leaving aside the issue of the leakage effects of trade and environmental policies.

The remainder of this chapter is organised as follows. In the first section, we describe the economic and political features of the model. In the second section, we characterise unilateral and cooperative equilibrium policies for the case of two symmetric countries. The third section analyses the relationship between green and producer interests. The final section briefly presents some concluding remarks.

The model

The economy

We consider two large countries, denominated home and foreign (denoted by *). Our analysis is focused on the economic and political structure of the home country (the foreign country will have symmetric characteristics).

The economy is described by a Ricardo–Viner model in which there are $N+1$ goods $i = 0, 1, \ldots, N$. All goods are produced under constant returns to scale and sold in a competitive market. Production of the numeraire good 0 requires labour alone and does not generate pollution. Production of all other goods requires both the mobile factor, labour, and a sector specific capital, and generates emissions at the fixed rate β per unit of output.

The numeraire good is traded freely across countries, with a world and domestic price equal to one. In a competitive equilibrium, this implies that wage rate is also equal to unity.

Let π_i be the international price of a non-numeraire good and q_i and p_i be its domestic consumer and producer prices, respectively. The reward to the owners of a specific factor can be denoted as $\Pi_i(p_i)$. By Hotelling's Lemma, the industry supply curve is then equal to $Y_i(p_i) = \partial\Pi_i/\partial p_i$, where $\partial Y_i/\partial p_i > 0$, and $\partial Y_i/\partial p_i^2 \leq 0$.

The economy is populated by H individuals, $h = 0, 1, \ldots, H$, with identical preferences. Utility is quasilinear and additively separable:

$$u_h(c_0, \ldots, c_N, Z) \equiv c_0 + \sum_{i=1}^{N} u_i(c_i) - Z, \tag{11.1}$$

where c_0 and c_i indicate consumption of the numeraire and non-numeraire goods. The functions $u(c_i)$ are differentiable, increasing, and strictly concave. The last term captures the disutility caused by environmental damage:

$$Z(\mathbf{p},\mathbf{p}^*) \equiv \sum_{i=1}^{N}\Big[(1-\theta_i)\beta_i Y_i(p_i)+\theta_i\beta_i^* Y_i^*(p_i^*)\Big], \tag{11.2}$$

where $\mathbf{p}$ and $\mathbf{p}^*$ are vectors of producer prices and $(1-\theta_i)$ and θ_i are the relative weights associated with domestic and foreign emissions in sector i, respectively. Equation (11.2) implies that, if the coefficient θ_i is positive, citizens in the home country are negatively affected by the emissions generated in both the domestic and foreign production of good i. The larger is θ_i, the larger is the impact of foreign pollution on the environmental damage suffered by the home citizens.

Inverse demand for a non-numeraire good can be expressed as a function of its price alone, i.e. $D_i(q_i)$. The indirect utility function corresponding to (11.1) can be written as:

$$\begin{aligned} V_h(\mathbf{q},\mathbf{p},\mathbf{p}^*) \quad &\equiv \quad L_h+\sum_{i=1}^{N}\lambda_i^h\Pi_i(p_i)+\frac{1}{H}\sum_{i=1}^{N}t_iY_i(p_i)+\frac{1}{H}\sum_{i=1}^{N}\tau_i\Big[D_i(q_i)-Y_i(p_i)\Big] \\ &+ \quad \sum_{i=1}^{N}u\Big(D_i(q_i)\Big)-\sum_{i=1}^{N}q_iD_i(q_i)-Z(\mathbf{p},\mathbf{p}^*). \end{aligned} \tag{11.3}$$

The terms in the first row of (11.3) represent income, which consists of wage income (L_h), capital claims (with λ_h indicating the share of capital owned by individual h)[6] and $1/H$ of environmental and trade revenues, transferred as a lump sum. The first two terms in the second row capture consumer surplus and the last term indicates environmental damage.

We consider two policy instruments: environmental taxes/subsidies $\mathbf{t}$ and import tariffs/subsidies τ. Thus the consumer prices of a non-numeraire good is given by $q_i=\pi_i+\tau_i$ and its producer price is $p_i=\pi_i+\tau_i-t_i$.

International product markets clear when

$$M_i(\pi_i,\tau_i,t_i)+M_i^*(\pi_i,\tau_i^*,t_i^*)=0, \qquad \forall i=1,\ldots,N, \tag{11.4}$$

where $M_i=D_i(q_i)-Y_i(p_i)$ and $M_i^*=D_i^*(q_i)^*-Y_i^*(p_i)^*$ represent the net imports of the home and foreign countries.

The leakage effects of trade and environmental policies

In the setup described above, both countries are 'large' in that they are able to affect world prices. In such a scenario, a unilateral increase in pollution taxes or a unilateral tariff cut will raise world prices and hence lead to an increase in foreign emissions. The indirect effects of domestic policies on foreign emissions via trade are referred to in the literature as *emission leakages*.

Formally, an increase in the domestic pollution tax on good i[7] generates the following effect on its international price:

$$\frac{\partial \pi}{\partial t} = -\frac{Y_p}{M' + M^{*\prime}} \equiv \delta, \tag{11.5}$$

where $M' = D_q - Y_p$, with $Y_p = \partial Y/\partial p$ and $D_q = \partial D/\partial q$. Notice that δ always lies between 0 and 1, implying an increase in the international price. Therefore, a unilateral increase in domestic pollution taxes shifts the comparative advantage of producing 'dirty' goods in favour of the foreign country and generates the following leakage effect:

$$\frac{\partial E^*}{\partial t} = \beta^* Y_p^* \delta. \tag{11.6}$$

Hence, if pollution taxes are raised unilaterally and unaccompanied by the use of import tariffs, they can only reduce domestic pollution at the cost of increased foreign pollution. It is important to stress that what is leaking through trade is not domestic pollution but domestic environmental policy. Thus, a leakage could also arise if environmental problems are strictly local. However, it is only in the case of transboundary environmental problems ($\theta_i > 0$) that the leakage negatively affects domestic residents. In this case, the environmental impact of an increase in the domestic pollution tax from the point of view of domestic residents is

$$\frac{\partial Z}{\partial t} = (1-\theta)\beta Y_p(\delta - 1) + \theta \beta^* Y_P^* \delta; \tag{11.7}$$

hence in the presence of trade flows higher domestic pollution taxes have two opposite environmental effects: a *direct positive effect*, due to a reduction in domestic emissions by $(1-\theta)\beta Y_p(\delta - 1)$; and an *indirect negative effect*, due to an increase in foreign transboundary emissions by $\theta \beta^* Y_p^* \delta$. The relative importance of the negative environmental effect increases with the size of the emission leakages and the degree to which foreign emissions cross over into the home country. Therefore,

Proposition 11.1 *A unilateral increase in pollution taxes, if unaccompanied by an increase in import tariffs, can lead to environmental degradation. A sufficient condition for this to occur is that the indirect environmental costs associated with the increase in transboundary foreign emissions outweigh the direct environmental benefits due to the reduction in domestic emissions.*

Consider now the impact of a unilateral increase in domestic import tariffs on the international price:

$$\frac{\partial \pi}{\partial \tau} = -\frac{M'}{M' + M^{*\prime}} \equiv -\phi. \tag{11.8}$$

Since $0 < \phi < 1$, higher domestic tariffs imply a fall in the international price and a shift of the terms of trade in favour of the implementing country. This generates the following leakage effects:

$$\frac{\partial E^*}{\partial \tau} = -\beta^* Y_p^* \phi. \tag{11.9}$$

The overall environmental impact of the domestic tariff increase is

$$\frac{\partial Z}{\partial \tau} = (1-\theta)\beta Y_p(1-\phi) - \theta\beta^* Y_p^* \phi. \tag{11.10}$$

Therefore, in the presence of trade flows higher domestic pollution taxes have two opposite environmental effects: a *positive effect*, due to a reduction in transboundary foreign emissions by $-\theta\beta^* Y_p^* \phi$; and a *negative effect*, due to an increase in domestic emissions by $(1-\theta)\beta Y_p(1-\phi)$. The relative importance of the negative environmental effect decreases with the size of the emission leakages and the degree to which foreign emissions cross over into the home country. The following result immediately arises:

Proposition 11.2 *A unilateral increase in import tariffs, if unaccompanied by an increase in pollution taxes, can lead to environmental degradation. A sufficient condition for this to occur is that the environmental costs associated with the increase in domestic emissions outweigh the environmental benefits due to the reduction in transboundary foreign emissions.*

To summarise the results obtained in this section, when emission taxes and import tariffs are selected unilaterally and are not combined, they can only reduce pollution in one country at the cost of increased pollution in the other country.

Propositions 11.1 and 11.2 will be key in understanding the relationship between green and producer lobbies. In Section 4, we will show that, in the presence of emission leakages such relationship will be ambiguous; if, however, the leakages effects of domestic policies are eliminated either through the combined use of pollution taxes and import tariffs or through international policy coordination, the relationship between the two lobbies will be unambiguous.

The political process

Our model does not explain the process of lobby formation. We simply assume that only the following groups of citizens can overcome the free-riding problem described by Olson (1965) and get politically organised: a proportion s^E of the population, the 'environmentalists', who form a national green lobby; and the owners of a subset S of all specific factors, who form producer lobbies in their respective sectors. In each sector $i \in S$, capital owners represent a proportion s^P of the population.

Political competition can be modelled as a two-stage game. In the first stage, green and producer lobbies simultaneously present incumbent policymakers with contribution schedules, namely functions mapping every combination or trade and environmental policy into a level of political contribution. We assume that a citizen cannot be a member of more than one interest group. We also exclude the possibility that lobbies cooperate with one another and that they can offer political contributions to politicians in the other country. Therefore, when we refer to an 'alliance' between green and producer lobbies, we will be alluding to the

fact that they exercise political pressure in the same direction, without formally coordinating their actions. The equilibrium set of contribution schedules is one in which each lobby maximises the aggregate utility of its members, given the schedules of the other lobby group.

In the second stage, incumbent politicians select trade and environmental policies, given the equilibrium contribution schedules, and collect the corresponding contributions from every lobby. They are concerned with aggregate well-being, but also with the support they get from interest groups. In equilibrium, the decision-makers balance optimally the marginal benefit of net aggregate contributions against the marginal welfare cost of distortionary trade and environmental policies.

In contrast to Grossman and Helpman (1994), we assume that interest groups are 'functionally specialised' (Aidt, 1998), in the sense that producer lobbies are only concerned about industry profits and the green lobby is only concerned about environmental damage. The gross (of contributions) welfare of a producer lobby $i \in S$ is thus given by:

$$W_i^P(t_i, \tau_i, t_i^*, \tau_i^*) \equiv s^P H \Pi_i(t_i, \tau_i, t_i^*, \tau_i^*), \qquad \forall i \in S, \tag{11.11}$$

while the utility of the national environmental lobby is

$$W^E(\mathbf{t}, \tau, \mathbf{t}^*, \tau^*) \equiv B - s^E H Z(\mathbf{t}, \tau, \mathbf{t}^*, \tau^*), \tag{11.12}$$

where B is a constant.

National producer and green lobbies present their government with contribution schedules $C_i(\mathbf{t}, \tau; \mathbf{t}^*, \tau^*)$. Their objective functions are, respectively,

$$\tilde{W}_i^P(\mathbf{t}, \tau, \mathbf{t}^*, \tau^*) \equiv W_i^P(t_i, \tau_i, t_i^*, \tau_i^*) - C_i(t_i, \tau_i; t_i^*, \tau_i^*), \qquad \forall i \in S, \tag{11.13}$$

and

$$\tilde{W}^E(\mathbf{t}, \tau, \mathbf{t}^*, \tau^*) \equiv W^E(\mathbf{t}, \tau, \mathbf{t}^*, \tau^*) - \sum_i C_i(t_i, \tau_i; t_i^*, \tau_i^*). \tag{11.14}$$

The implicit objective of incumbent politicians is to be reelected.[8] This implies that they care about the utility level achieved by the median voter, particularly if voters are well informed about the effects of government policy and base their vote partly on their standard of living. Incumbent politicians also value political contributions for financing future campaigns and deterring competitors. The government's objective is thus given by

$$G(\mathbf{t}, \tau, \mathbf{t}^*, \tau^*) \equiv aW(\mathbf{t}, \tau, \mathbf{t}^*, \tau^*) + \sum_i C_i(t_i, \tau_i; t_i^*, \tau_i^*), \qquad a \geq 0, \tag{11.15}$$

where W is the welfare of citizens (or 'social welfare') and a represents the weight that the government attaches to social welfare relative to lobbies' contributions.

Social welfare is defined as aggregate income plus total consumer surplus minus total environmental damage:

$$W(\mathbf{t},\tau,\mathbf{t}^*,\tau^*) \equiv L+\sum_{i=1}^{N}\Pi_i(t_i,\tau_i,t_i^*,\tau_i^*)+\sum_{i=1}^{N}t_iY_i(t_i,\tau_i,t_i^*,\tau_i^*)+\sum_{i=1}^{N}\tau_iM_i(t_i,\tau_i,t_i^*,\tau_i^*)$$
$$+ H\left[\sum_{i=1}^{N}u\Big(D_i(t_i,\tau_i,t_i^*,\tau_i^*)\Big)-\sum_{i=1}^{N}q_iD_i(t_i,\tau_i,t_i^*,\tau_i^*)\right]-HZ(\mathbf{t},\tau,\mathbf{t}^*,\tau^*). \tag{11.16}$$

In order to derive the equilibrium cooperative policies, we can rely on the notion that the outcomes of international negotiations must satisfy Pareto efficiency for the two policy-makers involved (see Grossman and Helpman, 1995). This implies that cooperative policies must maximise the weighted sum

$$G^W \equiv a^*G+aG^* \quad = a^*a\Big[W(\mathbf{t},\tau,\mathbf{t}^*,\tau^*)+W^*(\mathbf{t}^*,\tau^*,\mathbf{t},\tau)\Big]+$$
$$a^*\textstyle\sum_i C_i(t_i,\tau_i,t_i^*,\tau_i^*)+a\sum_i C_i^*(t_i^*,\tau_i^*,t_i,\tau_i). \tag{11.17}$$

Thus the cooperative equilibrium policies are the same that would be selected by a single decision (a 'supra-national mediator') with preferences as given on the right-hand side of (11.17)[9]

Common agency games of the types described typically admit a multiplicity of Nash equilibria. Following Grossman and Helpman (1994), we focus on *truthful equilibria*, where lobbies make contributions up to the point where the resulting change in economic policies is exactly offset by the marginal cost of the contributions.

The policy equilibria

In this section, we characterise the (politically) optimal unilateral and cooperative equilibrium policies in a sector $i \in S$ of the economy.[10] We focus on the simple case in which the two countries have identical economic and political structures and consider three alternative policy regimes: one where governments have control over both trade and environmental policies; one in which they are restrained to the use of environmental policy by an existing free trade agreement; and one in which trade policy is the only instrument at their disposal.

Trade and environmental outcomes

Let us first consider the case where governments set trade and environmental taxes independently. Using the equilibrium conditions given in Conconi (2001), we obtain:

$$\tau_{NC}=\tau_{NC}^*=\frac{\beta H\theta Y_p(a+s^E)}{a(Y_p-D_q)}, \tag{11.18}$$

and

$$t_{NC}=t_{NC}^*=\frac{H[\beta Y_p(a+s^E)(1-\theta)-s^PY]}{aY_p}. \tag{11.19}$$

In the case of centralised decision-making, governments select the following policies:

$$\tau_C = \tau_C^* = 0, \tag{11.20}$$

and

$$t_C = t_C^* = \frac{H[\beta Y_p(a+s^E) - s^P Y]}{aY_p}. \tag{11.21}$$

Environmental-only outcomes

Next, consider the case in which the two governments have signed a free trade agreement, eliminating the tariffs on each other's imports. In this scenario, environmental policy is the only instrument available. Unilateral emissions are given by

$$t_{NC} = t_{NC}^* = \frac{H[\beta Y_p(a+s^E)(\delta+\theta-1) - s^P Y(\delta-1)]}{aY_p(\delta-1)}, \tag{11.22}$$

while international policy coordination yields

$$t_C = t_C^* = \frac{H[\beta Y_p(a+s^E) - s^P Y]}{aY_p}. \tag{11.23}$$

Trade-only outcomes

Finally, suppose trade policy is the only instrument available. Unilateral policy-making leads to the adoption of the following import tariffs:

$$\tau_{NC} = \tau_{NC}^* = \frac{H[\beta Y_p(a+s^E)(1-2\theta) - s^P Y]}{a(D_q - Y_p)}, \tag{11.24}$$

while cooperative policy-making results in the adoption of identical import tariffs:

$$\tau_C = \tau_C^*. \tag{11.25}$$

Green and producer lobbies: competition or alliance?

In this section, we examine the impact of lobbying by green and producer groups on the policy outcomes derived above. This then allows us to evaluate whether green and producer lobbies have similar or divergent interests over trade and environmental policy. As a measure of a lobby's influence, we consider the effect of a change in its size on the policy outcomes, i.e. $\partial\tau/\partial s^E$ and $\partial t/\partial s^E$ for the green lobbies and $\partial\tau/\partial s^P$ and $\partial t/\partial s^P$ for producer lobbies.

Let us examine each of the policy scenarios considered in the previous section, starting from the case in which governments can use both policy instruments and act in a non-cooperative manner. We obtain the following result:

Lemma 11.1 *If two symmetric governments select trade and environmental policies unilaterally, green and producer lobbies will have opposite interests over environmental policy.*

Proof: Green lobbying leads to an increase in the pollution tax by

$$\frac{\partial t_{NC}}{\partial s^E} = \frac{\beta H(1-\theta)}{a} > 0, \tag{11.26}$$

and to an increase in the import tariff by

$$\frac{\partial \tau_{NC}}{\partial s^E} = \frac{\beta H \theta Y_p}{a(Yp - Dq)} > 0. \tag{11.27}$$

Producer lobbying leads to a fall in the pollution tax by

$$\frac{\partial t_{NC}}{\partial s^P} = -\frac{HY}{aY_p} < 0, \tag{11.28}$$

and has no effect on equilibrium import tariffs:

$$\frac{\partial \tau_{NC}}{\partial s^P} = 0. \tag{11.29}$$

Q.E.D.

Moving to the case of centralised decision-making, we find:

Lemma 11.2 *If two symmetric governments select trade and environmental policies cooperatively, green and producer lobbies will have opposite interests over environmental policy.*

Proof: The presence of the green lobby implies an increase in cooperative pollution taxes:

$$\frac{\partial t_C}{\partial s^E} = \frac{\beta H}{a} > 0. \tag{11.30}$$

The impact of producer lobbying on the cooperative equilibrium policies is

$$\frac{\partial t_C}{\partial s^P} = -\frac{HY}{aY_p} < 0. \tag{11.31}$$

None of the lobbies has any impact on the trade policy outcomes:

$$\frac{\partial \tau_C}{\partial s^E} = \frac{\partial \tau_C}{\partial s^P} = 0. \tag{11.32}$$

Q.E.D.

Consider now the situation in which governments have committed to free trade. In the case of decentralised decision-making, we obtain the following result:

Lemma 11.3 *If two symmetric governments select emission taxes non-cooperatively, green and producer lobbies will have opposite interests over environmental policy if and only if* $\delta + \theta < 1$.

Proof: Under a free trade regime, green lobbying has an ambiguous effect on the non-cooperative environmental outcomes:

$$\frac{\partial t_{NC}}{\partial s^E} = \frac{\beta H(\delta + \theta - 1)}{a(\delta - 1)}. \tag{11.33}$$

It is straightforward to verify that expression (11.33) is positive for $\delta + \theta < 1$. This condition implies green lobbies will support a unilateral increase in pollution taxes only if the environmental benefits associated with the decrease in domestic pollution outweigh the environmental costs due to the increase in foreign transboundary pollution. The impact of producer lobbying on the unilateral environmental policy outcomes is

$$\frac{\partial t_{NC}}{\partial s^P} = -\frac{HY}{aY_p} < 0. \tag{11.34}$$

Q.E.D.

If the decision-making process is centralised, the relationship between environmental and producer groups is described by the following lemma:

Lemma 11.4 *If two symmetric governments select emission taxes cooperatively, green and producer lobbies will always have opposite interests over environmental policy.*

Proof: Green lobbying biases cooperative emission taxes upwards:

$$\frac{\partial t_C}{\partial s^E} = \frac{\beta H}{a} > 0, \tag{11.35}$$

while producer lobbying has the opposite effect:

$$\frac{\partial t_C}{\partial s^P} = -\frac{HY}{aY_p} < 0. \tag{11.36}$$

The competitive nature of the relationship between the two lobbies is due to the fact that a multilateral increase in emission taxes will unambiguously lead to a reduction in productive activities in both countries, which implies a reduction in total environmental damage and a fall in industry profits in both countries. Q.E.D.

Let us now consider the scenario in which trade policy is the only instrument available. When import tariffs are selected in an independent manner, we obtain:

Lemma 11.5 *If two symmetric governments select import tariffs unilaterally, green and producer lobbies will have opposite interests if and only if* $\theta < 1/2$.

Proof: An increase in the size of the green lobby has the following impact on non-cooperative import tariffs:

$$\frac{\partial \tau_{NC}}{\partial s^E} = \frac{\beta H Y_p (1-2\theta)}{a(D_q - Y_p)}, \tag{11.37}$$

which is negative for $\theta < 1/2$. This implies that in the case of local or regional environmental problems ($\theta \leq 1/2$) green lobbying will bias import tariffs downwards, since in this case the environmental costs associated with the increase in domestic emissions will outweigh the environmental benefits due to the fall in foreign emissions; in the case of global environmental problems ($\theta = 1/2$), green lobbying will have no effect on the trade policy outcomes, since the environmental gains associated with the decrease in foreign pollution will exactly offset the costs associated with the increase in domestic emissions. The impact of producer lobbying is:

$$\frac{\partial \tau_{NC}}{\partial s^P} = \frac{HY}{a(Y_p - D_p)} > 0. \tag{11.38}$$

Q.E.D.

Finally, Lemma 11.6 applies to the case of trade policy coordination:

Lemma 11.6 *If two symmetric governments select import tariffs cooperatively, green and producer lobbies have no impact on the policy outcomes.*

Proof: an increase in the side of the green or producer lobby has no effect on the cooperative equilibrium tariffs:

$$\frac{\partial \tau_C}{\partial s^E} = \frac{\partial \tau_C}{\partial s^P} = 0. \tag{11.39}$$

This result is due to the fact that in equilibrium two symmetric countries will always adopt identical import tariffs and trade policy will thus have no impact on productive activities and emission levels. Q.E.D.

The results presented in Lemmas 11.1—11.6 are summarised by Table 11.1 and by the following Proposition:

Proposition 11.3 *The nature of the relationship between green and producer lobbies depends crucially on which policy instruments are available, whether government act in a unilateral or cooperative manner, and the magnitude of the emission leakages and the associated transboundary spillovers.*

Table 11.1 shows that the ambiguity of the relationship between green and producer groups arises only in the presence of emission leakages (cases 3 and 5). If instead emission leakages are eliminated either through the combined use of trade and environmental policy (cases 1 and 2) or through international policy coordination (cases 2, 4, and 6), the relationship between green and producer groups is always unambiguous.

Table 11.1 The relationships between green and producer lobbies

Policy regimes	*Policy-making process*	
	Decentralised	*Centralised*
Trade and environment	1 Competition over environmental policy	2 Competition over environmental policy
Environment only	3 Competition over environmental policy if $\delta + \theta < 1$	4 Competition over environmental policy
Trade only	5 Competition over trade policy if $\theta < 1/2$	6 —

Concluding remarks

In this chapter we have employed a common agency model to examine the role of green and producer lobbies in the joint determination of trade and environmental policy. We have focused our analysis on the case of two large symmetric countries, which are linked through trade and transboundary pollution.

We have characterised the policy outcomes and the relationship between lobbies in three alternative policy regimes: one where governments control both trade and environmental policies; one in which they are restrained to the use of environmental policy by an existing free trade agreement; and one in which trade policy is the only available instrument.

We have shown that, when domestic policy generate emission leakages, the relationship between green and producer interests over trade and environmental policy is ambiguous. If the emission leakages are eliminated through the combined use of trade and environmental policy instruments or through international policy cooperation, green and producer lobbies will unambiguously be *either* enemies or allies.

The model presented in this chapter is characterised by the existence of three types of distortions: an environmental distortion, caused by the presence of emission spillovers; a trade distortion, due to the fact that countries are able to affect the terms of trade; and a political distortion, arising from the lobbying activities of green and producer groups. Is it still possible to achieve efficient policy outcomes in this second-best world? The first-best solution, which is obtained when benevolent policymakers act cooperatively, requires that governments eliminate tariffs on each other's imports and adopt optimal Pigouvian emission taxes, which reflects the social marginal damage of emissions. In the presence of green and producer

lobbies, cooperative environmental policies will be efficient only if green lobbies are large enough to offset the political pressure exercised by producer lobbies. For unilateral environmental policies to be efficient, green lobbies must be even larger, so that their bias towards higher taxes counteracts the downward bias of both producer groups and the national governments. The United States's defection from a cooperative agreement like the Kyoto Protocol, in the face of a strong producer lobby, suggests that efficiency can only be achieved through a larger domestic green lobby.

Acknowledgements

I wish to thank Brian Copeland, Carlo Perroni, Alistair Ulph, John Whalley and Myrna Wooders for comments and suggestions. All errors are mine. This research was supported by a University of Warwick Graduate Award and an ESRC studentship.

Notes

1 The North American Agreement on Environmental Cooperation (NAAEC), could be characterised as being primarily concerned with safeguarding the sovereign rights of each party to establish its environmental standards while working towards the compatibility of standards.
2 Opposition on the part of business and environmental groups has also undermined the project of a Free-Trade Area of the Americas (FTAA), which the United States, Canada and 34 American and Caribbean countries (all of them except Cuba) have agreed to establish by 2005.
3 See *The Economist*, 11 December 1999.
4 In this area, the political contributions approach developed by Grossman and Helpman (1994, 1995, 1996) has become something of a work-horse model. For an extensive review of this literature, see Persson and Tabellini (2000). See also, Rama and Tabellini (1998); Mitra (1999); and Gawande and Bandyopadhyay (2000).
5 Hillman and Ursprung (1988, 1994) investigate how environmental concerns might affect international trade policy. Fredriksson (1997) and Aidt (1998) examine the effect of lobbying by green and producer groups on the determination of environmental policy.
6 We assume that individuals own at most one type of specific factor.
7 Given the quasilinearity of the utility function, there is no possibility of substitution among goods such that the amount of pollution resulting from a given level of production can be varied. This allows us to study the determination trade and environmental policies in a representative non-numeraire sector i of the economy. For ease of the exposition, in what follows we drop the sectoral subscript.
8 See Grossman and Helpman (1996) for an explicit treatment of the electoral stage.
9 Notice that (11.17) stipulates that cooperative policies must be efficient for the two governments without specifying how the surplus will be divided between them. To determine which utility pair (G, G^*) will be selected, a bargaining procedure should be introduced. One could adopt the Nash bargaining solution, or the Rubinstein's bargaining solution.
10 The equilibrium conditions for unilateral and cooperative trade and environmental policies can be found in Conconi (2001).

References

Aidt, T.S. (1998) 'Political internalization of economic externalities and environmental policy', *Journal of Public Economics* 69: 1–16.

Bernheim, B.D. and M.D. Whinston (1986) 'Menu auctions, resource allocation, and economic influence', *Quarterly Journal of Economics*, 101: 1–31.

Conconi, P. (2001) 'Green and producer lobbies: competition or alliance?', Warwick University CSGR Working Paper no. 78/01.

Conconi, P. (2002) 'Green lobbies and transboundary pollutionin large open economies', *Journal of International Economics*, forthcoming.

Fredriksson, P. (1997) 'The political economy of pollution taxes in a small open economy', *Journal of Environmental Economics and Management*, 33: 44–58.

Fredriksson, P. (1999) 'The political economy of trade liberalization and environmental policy', *Southern Economic Journal*, 65: 513–25.

Gawande, K. and U. Bandyopadhyay (2000) 'Is protection for sale? Evidence on the Grossman-Helpman theory of endogenous protection,' *Review of Economics and Statistics*, 82(1): 139–52.

Grossman, G.M. and E. Helpman (1994) 'Protection for sale', *American Economic Review*, 84: 833–50.

Grossman, G.M. and E. Helpman (1995) 'Trade wars and trade talks', *Journal of Political Economy*, 103: 675–708.

Grossman, G.M. and E. Helpman (1996) 'Electoral competition and special interest politics', *Review of Economic Studies*, 63: 265–86.

Hillman, A.L. and H.W. Ursprung (1988) 'The influence of environmental concerns on the political determination of trade policy', in K. Anderson, and R. Blackhurst (eds) *The Greening of World Trade Issues*, New York: Harvester Wheatcheaf.

Hillman, A.L. and H.W. Ursprung (1994) 'Greens, supergreens, and international trade policy: environmental concerns and protectionism', in C. Carraro (ed.) *Trade, Innovation, Environment*, Dordrecht: Kluwer Academic.

Husted, A.L. and J.M. Logsdon (1997) 'The impacts of NAFTA on Mexico's environmental policy', *Growth and Change*, 28: 24–48.

Mitra, D. (1999) 'Endogenous lobby formation and endogenous protection: a long-run model of trade policy determination', *American Economic Review*, 89: 1116–34.

Olson, M. (1965) *The Logic of Collective Action*, Cambridge, MA: Harvard University Press.

Perroni, C. and R.M. Wigle (1994) 'International trade and environmental quality: how important are the linkages?', *Canadian Journal of Economics* 27: 551–67.

Persson, T. and G. Tabellini (2000) *Political Economics and Economic Policy*, Cambridge, MA: MIT Press.

Rama, M. and G. Tabellini (1998) 'Lobbying by capital and labor over trade and labor market policies', *European Economic Review*, 42: 1295–1316.

Schleich, J. (1999) 'Environmental quality with endogenous domestic and trade policies', *European Journal of Political Economy* 15: 53–71.

VanGrasstek, C. (1992) 'The political economy of trade and environment in the United States', in P. Low, *International Trade and the Environment*, Washington, DC: World Bank.

12 Income tax enforcement with a self-interested auditor

Martin Summer and Muriel Niederle

Income tax policy in most countries is implemented and enforced by a procedure that requires taxpayers to report their income in a preliminary round of information transmission. Based on this information some reports are then audited and penalties are applied to detected evaders. The tax compliance problem has attracted much academic attention. In particular, optimal mixtures of tax policy, auditing probabilities and penalties have been studied. Lots of useful insights about the economics of enforcement emerged from this literature (see Mookherjee, 1989; Andreoni *et al.* 1998). Perhaps somewhat surprisingly most authors usually abstract from the fact that in reality auditing has to be done by employees of the enforcement authority who might not automatically share its goals. Usually auditing activities are modelled by an impersonal auditing technology, which can be operated at some cost by the enforcer in line with his objectives.

In this chapter we study a model that takes into account that auditing activities are delegated to employees of the enforcement authority. Since they have nothing to do with the organisation's goals *per se* it is unlikely that they will just deliver detection of non-compliers in the same way as an impersonal technology. Since the results of the audit will depend on whether an auditor worked hard or not and since there is hardly any practical way to observe whether this was indeed the case he might minimise his effort on auditing by just rubberstamping most of the files assigned for investigation, if he is not appropriately motivated by incentives. In the jargon of modern information economics: most practical enforcement problems will have to face some form of auditor moral hazard. The problem of motivating auditors is thus an issue of considerable practical importance for an effective organisation of any real world enforcement process.

Our chapter makes a first step to analyse this issue. In particular it makes two main contributions. First, we propose a game theoretic framework to describe auditor moral hazard in the process of income tax enforcement. Second, we show that the nature of the problem imposes quite narrow limits on the possibilities for the traditional economic resolution of incentive problems by a system of monetary rewards and punishments.

The central insight of our analysis is that a simple incentive wage scheme for the auditor that rewards detection of evaders and punishes investigations of files that afterwards turn out to be unjustified is likely to lead to poor performance of the enforcement procedure as it is not just the auditor's incentives that matter. The incentives of the auditor and the taxpayer are interrelated. The enforcement authority cannot easily manipulate the whole *system* of incentives that is decisive for the performance of the enforcement process. Important parameters, which affect the taxpayer's incentives as the tax code and the legal consequences of evasion, are usually given to the enforcer. The game theoretic analysis of the strategic interaction between an auditor and taxpayers also shows that for a large set of parameter values there are equilibria where an incentive wage schedule works to resolve the problem of auditor moral hazard but still he is ineffective as an enforcement instrument. Thus there are good theoretical reasons to expect narrow intrinsic limits to an incentive contract solution to the problem of turning a self-interested auditor into an *effective* enforcement instrument. Monetary incentives for the auditor might not be very powerful.

We therefore argue that academic interest in the tax enforcement problem should perhaps concentrate more intensively on the question of how the potential of *intrinsic motivation* on the part of the auditor *and* the taxpayers can be systematically supported by appropriate rules and institutions. These aspects of the delicate interplay between intrinsic and extrinsic motivation have been widely discussed among social psychologists and sociologists but are largely neglected in the mainstream economic literature.[1] Though our model does not directly contribute to the analysis of intrinsic motivation and non-monetary incentives in income tax policy enforcement, by pointing out the limits of enforcement relying exclusively on tools of extrinsic motivation – like incentive payments fines and monitoring – it contributes to the argument that these other sources of motivation and incentives should be more thoroughly studied in the context of income tax enforcement.

There are two closely-related problems to which our analysis might apply in an analogous way. First, the control problem arising from public sector programmes which provide financial assistance such as grants and loans to persons or institutions. The applications are frequently based on claims about imperfectly – or even unobservable – characteristics. Secondly, within modern firms the internal supervision of divisions to safeguard assets and records to protect them from being stolen, misused or destroyed, is a related problem. It should be pointed out, however, that the spirit of our model as well as some of the results show that institutional details are crucial for the understanding of practical enforcement problems. Our model is constructed with tax enforcement in mind. The problems mentioned above might require additional features and – though related – should not just be simply subsumed under our framework. Nevertheless we believe that our model provides some relevant insights for these practical problems as well.

The rest of the chapter is organised as follows. The first section presents an overview of the model. The second section is concerned with related research. The third section describes the model and discusses our various assumptions. The fourth section is about beliefs, optimal inference and the auditor's response and

the fifth section considers boundary cases. The sixth section is about equilibrium outcomes, and the final section concludes. Proofs are in an appendix.

An overview of the model

Our model is focused on the analysis of how an auditor's and a taxpayer's incentives are interrelated.[2] We describe the auditing process as an extensive form game with imperfect information between an auditor and a taxpayer where the tax authorities or the internal revenue service (IRS) is lurking in the background and enters the model explicitly only in form of payoff parameters. Taxpayers are characterised by their income, which is private information and for simplicity is assumed to fall into two categories. The income can be high or low. Only the distribution of income is common knowledge. Taxpayers have to report their income to the IRS, and an auditor investigates the files. There is a progressive tax schedule and the high-income type has an incentive to report a low income. Based on the report the auditor has to decide whether he should conduct an investigation or not. This is the moral hazard part of the game. Since effort is assumed to be unobservable he can always do nothing and claim that he found no evidence of non-compliance. However, there is a simple (exogenous) incentive wage scheme to motivate the auditor to conduct proper audits. The incentive scheme rewards detection of evaders and punishes 'false alarms' or wrong accusations – whereas the clearing of a report leads to an intermediary payment. The agent has to pay a penalty if evasion is discovered but has an advantage if he gets away undetected.

To incorporate some realistic features of the enforcement process we postulate that the auditor can only discover conflicting evidence through an imperfect technology. Contradictions between the auditor and the taxpayer's report lead to a legal procedure in which the case is settled. So at the end of the day, either (i) a file is cleared, (ii) an evader is detected and punished (by a fine), or (iii) it turns out that a taxpayer was wrongly accused of evasion. Given the exogenous payoff structure we concentrate on the strategic interaction between the auditor and the agent and analyse sequential equilibria of this game depending on several interesting constellations of the parameters, which are the incentive scheme for the auditor, the legal system (described by the level of fines applied to convicted evaders), the tax system and the (imperfect) auditing technology.

Related research

There is an extensive literature on the economics of tax policy enforcement. For a survey of basic models and arguments we refer to the papers by Mookherjee (1989) and Andreoni *et al.* (1998). Our chapter builds on the literature on game theoretic models of auditing, where commitment to an auditing rule is not possible. Seminal papers in this literature are Reinganum and Wilde (1986) and Graetz *et al.* (1986). However, the central aspect of our model deals with auditor moral hazard in the process of enforcement. This issue has not been studied before. By placing auditor moral hazard at the centre of the analysis our chapter fundamentally differs

from other models of delegated enforcement as Melumad and Mookherjee (1989) or Sanchez and Sobel (1993), where the auditing process is represented by some impersonal detection technology.

Our framework has some parallels with a paper from the accounting literature by Baiman *et al.* (1987). These authors study a self-interested auditor problem. For mainly two reasons, their model cannot directly be applied to the income tax enforcement problem. First of all they assume that a principal can choose – among other things – a transfer scheme (which would correspond to a tax policy in our model) which in a descriptive theory of income tax enforcement should not be assumed. Tax policies are decided at the level of government authorities and not at the level of the enforcement authority. Second – as in other principal–agent models – their model requires the assumption of extensive commitment possibilities for the principal. These commitment possibilities are mitigated in important ways by the structure of tax enforcement problems and should also not be assumed in a descriptive theory (see Mookherjee (1989) for a detailed argument).

Our approach to modelling the auditing technology builds on contributions from safeguard systems analysis (Avenhaus *et al.* 1995), game theory (Green, 1990) and from the accounting literature (Shibano, 1990). The main direction of analysis in these papers is to study the connection between equilibrium strategies and optimal statistical tests for non-compliance with no role for problems of auditor moral hazard.

The model

We consider a population of taxpayers characterised by their income. For simplicity we assume that there are two income categories. A proportion λ has a low income θ_1 whereas the share of high incomes θ_2 is $1 - \lambda$. Income is a taxpayer's private information. Following the literature on games with incomplete information, we can interpret the proportions as a representative taxpayer's type chosen by an initial move of nature with probability λ or $1 - \lambda$. We will therefore refer to the population as the taxpayer of type $\theta_1(\theta_2)$. Taxpayers have to file an income tax report. We denote reported income – which may be different from the true income – by $\hat{\theta}_1(\hat{\theta}_2)$. For the other players taxpayers can not be distinguished *a priori*. In real world situations taxpayers might of course be distinguished by other criteria than income from data like age, profession, area of residence, usually known by the authorities. So our modelling choice can be best thought of as describing a class of taxpayers rather than providing a global description of the entire population.

Since a feature of real world tax policy enforcement is that the enforcement process is delegated, important parameters of relevance are already given. The tax policy specifies a progressive income tax. To enforce the tax policy, some reports are audited. We assume that an auditor is in charge of performing this task. The auditor investigates reports with an imperfect auditing technology and files his findings with the IRS. It is assumed that he can always claim to have found no evidence of non-compliance. However, it is not possible to falsely claim that an honest taxpayer has underreported. This possibility allows the auditor first to decide

whether he wants to conduct a proper investigation L for 'look' or just shirk D for 'don't look'. If he conducts a proper audit (when the auditor has put in effort) this yields some evidence or some *signal* x in ∇. The signal is a random variable and its distribution G depends on the true state. We denote by $G(x|\theta_i)$ the distribution function of x, given state θ_i, where $g(.|\theta_i)$ is its density for $i = 1,2$. After having observed the signal the auditor can take two actions. Either he can clear C for 'clear' the file, or, if he got conflicting evidence and wishes to reveal this information to the IRS, he can point out the irregularities (W for 'warn'). Based on self-reporting of the taxpayer and the reported results of the auditor either the case is cleared or there is a contradiction between the two reports. This leads to a legal investigation in which the true state is found out with certainty. The reason why in this case the legal procedure is not always used rightaway is based on an assumption. We assume that for the IRS the way to the courts is rather expensive (not explicitly modelled), and it really wants to take only a promising selection of cases to the courts. We think of the combination of an auditor and a legal procedure as if the IRS can enforce the given tax policy by a combination of an expensive perfect and a cheaper but imperfect auditing technology. Of course an administrative penalty or a court is not literally an auditing technology. However, at an abstract level the essential aspect of its function in the enforcement process can be adequately described by our approach, see Koffman and Lawaree (1993).

In our model payoffs are *exogenous*. In particular the payment schedule of the auditor is chosen such that he is maximally rewarded if his work leads to a successful conviction of an evader. He gets minimally rewarded when his report leads to a legal procedure but the taxpayer turns out to be innocent and he gets an intermediary payment for clearing a report. Though postulated exogenously this choice of payoffs for the auditor is not completely arbitrary in the context of our problem. One way how such a schedule can come about is when it could be chosen by the IRS with the objective of maximising net tax revenues. This is often postulated in the literature on tax policy enforcement. As a government authority the IRS acts in a legal environment that prescribes that enforcement activities have the function of general and specific deterrence. In the light of these principles net tax revenue maximisation is an adequate goal for the IRS. In order to achieve this goal with a self-interested auditor the IRS has to motivate him appropriately to function as an effective enforcement instrument.

To be specific about the payoffs, we normalise the payoff of the auditor and of both types (separately) such that in a situation where the taxpayer reports truthfully and the auditor doesn't conduct an investigation (the *status quo*) their payoffs are both 0.

Since the IRS can not distinguish a truthful report from a lie without the results from an audit the payoff to the auditor is always the *status quo* payoff if he doesn't investigate. Clearly the situation is not the same for the agent. We assume that the high type lying without being checked receives his highest payoff, which we normalise to one and the low type agent misrepresenting his type and not being checked receives a negative payoff normalised to minus one.

Now assume an investigation takes place. If the auditor clears a true report this is the same for the agent as the status quo. For the auditor, on the other hand, some disutility of effort $-e$ has to be incurred if he investigates. We assume $0 < e$. If $e \geq 0$ the auditor will always investigate. If the auditor wrongly accuses the taxpayer of non compliance the incentive scheme assigns to him the lowest possible payoff $-f$ (f for 'false alarm'). We assume that $0 < e < f$. If $f \leq e$ then the auditor after the investigation will always raise the alarm. Abstracting from the anxiety most people experience when they are audited, we assume that the taxpayer receives a payoff of zero as in the *status quo*.

If the auditor detects an overstatement he is paid just enough to get utility zero. (A positive payoff instead of zero would not change the results.) The taxpayer is assumed to have a disutility from such a deviation from truthful reporting and receives $-i$ (i for 'inconvenience'), where $0 < i$, which expresses the idea that for the taxpayer the inconvenience in case of a detected overstatement are strictly greater than the costs of a 'false alarm'. This inconvenience can arise from the time burden of correspondences, the necessity of meeting tax authorities or even to hire professional assistance, in short: all the trouble that can arise from such a form of 'misunderstanding'.

Detection of an understatement yields a reward for the auditor from which we have to deduct inspection costs. We denote his net reward as r. If the taxpayer's lie is found out he is punished and gets $-p$ (p for 'punishment'). We furthermore don't need to compare i to p. If the taxpayer lies and stays undetected he gets away with a payoff b (the 'booty'). To summarise, the institutional restrictions coming from the tax code and the legal system are described by the parameters b and p. The auditor's incentive scheme is captured by the parameters $r, 0$ and f. Our results do not depend on our normalisation, since for each player and each type only two payoffs are fixed.

Beliefs, optimal inference and auditor's best reply

We now want to study the equilibrium interactions resulting form this strategic situation. Note first that declaring a high income $\hat{\theta}_2$ is a strictly dominated strategy for the low income taxpayer. Therefore, whenever the auditor gets a report of a high income, he can conclude that this report is actually true and thus won't investigate such reports.[3] Let us denote by σ the probability that a high income type reports truthfully. Let τ denote the probability that the auditor properly investigates a low income report and δ the probability that the auditor will report an irregularity (choose W) when he receives a low income report and the signal is x.

We have to determine the equilibrium values of σ, τ and δ. Since the auditor has to take his decisions under imperfect information we have to determine his beliefs. A low report $\hat{\theta}_1$ could in principle come from both types. The auditor forms beliefs according to Bayes' rule. So if he is confronted with a low report his beliefs of actually facing a low type are given as follows: For all σ:

$$h(\theta_1|\sigma) = \frac{\lambda}{\lambda + (1-\lambda)(1-\sigma)} \tag{12.1}$$

$$h(\theta_2|\sigma) = \frac{(1-\lambda)}{\lambda + (1-\lambda)(\ -\sigma)} \tag{12.2}$$

The belief $h(\theta_1|\sigma)$ that the taxpayer is actually a low type is increasing in the probability σ with which the high-type taxpayer tells the truth. If the high type lies with probability one the belief of facing a low-type taxpayer is exactly equal to the *a priori* probability λ of a low type. In a similar way it can be easily checked that the beliefs $h(\theta_2|\sigma)$ also have the same intuitive properties.

Based on these a priori beliefs the auditor has to decide whether he audits or not. Not conducting an audit gives him the *status quo* payoff zero. If he conducts an investigation his payoff is determined as follows:

He receives a signal x. This signal can be used to update his beliefs to $H(\theta_i|x,\sigma)$ over the true type $\theta_i \in \Theta$ according to Bayes' rule, that is for all signals x and for every $\sigma \in [0,1]$:

$$H(\theta_i|x,\sigma) = \frac{h(\theta_i|\sigma)g(x|\theta_i)}{\sum_{j=1}^{2} h(\theta_j|\sigma)g(x|\theta_j)} \qquad i = 1,2 \tag{12.3}$$

or substituting equations (12.1) and (12.2) for $h(.)$:

$$H(\theta_1|x,\sigma) = \frac{\lambda g(x|\theta_1)}{\lambda g(x|\theta_1) + (1-\lambda)(1-\sigma)g(x|\theta_2)} \tag{12.4}$$

$$H(\theta_2|x,\sigma) = \frac{(1-\lambda)(1-\sigma)g(x|\theta_1)}{\lambda g(x|\theta_1) + (1-\lambda)(1-\sigma)g(x|\theta_2)} \tag{12.5}$$

Given these posterior beliefs about the true type the auditor has to decide whether he clears 'C' or warns 'W'.

His strategy 'W' strictly dominates 'C' if and only if under the given beliefs 'W' yields a strictly larger payoff:

$$H(\theta_1|x,\sigma)(-f) + H(\theta_2|x,\sigma)r > H(\theta_1|x,\sigma)(-e) + H(\theta_2|x,\sigma)(-e) \tag{12.6}$$

Thus the auditor will choose his strategy 'W' if and only if

$$\frac{H(\theta_2|x,\sigma)}{H(\theta_1|x,\sigma)} > \frac{f-e}{r+e} \tag{12.7}$$

Clearly the attractiveness of warning depends on the relation of the costs that arise from a wrong accusation to the benefits forgone if a lie remains undetected and the likelihood ratio of finding out the true type conditional on the signal and the reporting strategy of the taxpayer. These two ratios are driving the economics of the auditing problem. The next paragraph shows that our modelling choice of using a continuous rather than a discrete signal simplifies the model because it

allows us to characterise a sharp decision rule for the auditor and helps us to avoid tedious case distinctions.

Without loss of generality we assume that the signals x are ordered such that a 'higher' x refer to the high type taxpayer whereas a 'lower' x refer to the low type taxpayer, more precisely such that $\frac{H(\theta_2|x,\sigma)}{H(\theta_1|x,\sigma)}$ is increasing in x for a given probability of lying σ. If this is not the case this monotonicity can be achieved by a reordering of the signals. Furthermore, we also assume without loss of generality that we do not have two distinct signals that lead to the same a *posteriori* beliefs, since the auditor will never distinguish between them.

Therefore $\frac{H(\theta_2|x,\sigma)}{H(\theta_1|x,\sigma)}$ is strictly increasing in x for a given σ. In fact – as we would intuitively expect – this ordering of the signals is independent of σ, the probability that the high type taxpayer reports the truth.

The following assumption is useful for the analysis of the model.

Assumption: $\frac{H(\theta_2|x,\sigma)}{H(\theta_1|x,\sigma)}$ is continuously increasing in x and continuously decreasing in σ.

Now let us return to the auditor's decision problem: We already know that high signals refer to the high type taxpayer and the auditor receiving signal x strictly prefers to warn if and only if condition (12.7) holds. Now let

$$x_\sigma = \max\left\{x : \frac{H(\theta_2|x,\sigma)}{H(\theta_1|x,\sigma)} \leq \frac{f-e}{r+e}\right\} = \max\left\{x : \frac{(1-\lambda)(1-\sigma)g(x|\theta_2)}{\lambda g(x|\theta_1)} \leq \frac{f-e}{r+e}\right\} \tag{12.8}$$

If $\frac{H(\theta_2|x,\sigma)}{H(\theta_1|x,\sigma)} > \frac{f-e}{r+e}$ for all x we denote $x_\sigma = y$ for a signal that is not in the set of signals where he clears.

Then the auditor's best replies are easy to characterise. For all signals $x < x_\sigma$ the auditor will strictly prefer to clear whereas he will strictly prefer to warn if the signal $x > x_\sigma$. For x_σ he is indifferent between her two possible actions.

By our assumption x_σ is increasing in σ. Thus, the bigger σ, i.e. the higher the probability that the high type taxpayer tells the truth, the greater is x_σ, i.e. the bigger will be the willingness of the auditor to clear signals. Alternatively, a larger σ means that the auditor needs more evidence of a 'lie' to issue a warning. This cut-off-value x_σ is continuous in σ. Furthermore it is easy to show that x_σ is strictly increasing in σas long as x_σ is not equal to y or the greatest possible signal.

Some boundary cases

We first show how the auditor's payoff varies in σ, the probability that a high-income taxpayer reports his income truthfully. We will need this result later for the characterisation of equilibrium. Furthermore Lemma 12.1 gives us some insights how the incentive scheme for the auditor works in this model. The Lemma establishes a monotonicity property for the payoff of an auditor who has chosen to conduct a proper investigation (i.e., who has chosen L):

Lemma 12.1 *Suppose the taxpayer has chosen to play strategy σ. For an optimal choice x_σ the maximum expected payoff of the auditor when he chooses to look is:*

$$I(\sigma, x_\sigma|L) = \frac{\lambda}{\lambda+(1-\lambda)(1-\sigma)}E(x_\sigma) + \frac{(1-\lambda)(1-\sigma)}{\lambda+(1-\lambda)(1-\sigma)}F(x_\sigma)$$

Where

$$E(x_\sigma) = G(x_\sigma|\theta_1)(-e) + (1-G(x_\sigma|\theta_1))(-f)$$
$$F(x_\sigma) = G(x_\sigma|\theta_2)(-e) + (1-G(x_\sigma|\theta_2))r$$

If there exists σ^* *such that* $I(\sigma^*, x_{\sigma^*}) = 0$ *then* $I(\sigma, x_\sigma) > 0$ *for all* $\sigma < \sigma^*$ *and* $I(\sigma, x_\sigma) < 0$ *for all* $\sigma > \sigma^*$. *Furthermore* $I(\sigma, x_\sigma)$ *is strictly decreasing on* $\sigma \in \lfloor 0, \sigma^* \rfloor$

Proof: See Appendix.

To interpret this Lemma, observe that $I(\sigma, x_\sigma|L)$ denotes the payoff of an auditor who is investigating an income tax report using his best reply based on his beliefs. The incentive scheme functions such that from the perspective of the auditor it is relatively the best when a taxpayer lies. Of course the goal of such a scheme is the hope that this will deter evasion in equilibrium. This shows quite clearly that the incentives of the auditor and the taxpayer are related. Ultimately the IRS wants to achieve compliance. To accomplish this task with a self-interested auditor by monetary incentives, it must reward the auditor for detection of evaders. Therefore from the auditor's perspective evasion is good because it allows him to collect detection premia. The idea is to make him such a keen investigator that finally in *equilibrium* this deters potential evaders from cheating. So we see that the incentives of the taxpayer and the auditor *simultaneously* determine equilibrium behaviour.

The next result shows that there can in principle be a case where the auditor never audits despite the fact that the taxpayer chooses always to lie. This observation is obvious because it says that when the payoff from inspecting is lower than the *status quo* even in the relatively best case for the auditor it will certainly never pay for him to look. An implication of this observation is that *some* sort of motivation is necessary in any case to make a self-interested auditor effective in enforcement. Let us summarise this observation by:

Lemma 12.2 *If* $I(0, x_0|L) < 0$ *then the auditor will never audit at all and for all* $0 \leq \sigma \leq 1$ *will choose* $\tau = 0$

.

Assume the auditor audits with a certain probability τ and chooses an optimal cut-off signal x_σ. The taxpayer is willing to lie with probability $(1-\sigma)$ if and only if

$$(1-\tau)b + \tau[G(x_\sigma|\theta_2)b + (1-G(x_\sigma|\theta_2)(-p))] = 0 \tag{12.9}$$

which is equivalent to the equation

$$G(x_\sigma|\theta_2) = \frac{\tau p - (1-\tau)b}{\tau(b+b)} \tag{12.10}$$

In the case when $G(x_\sigma|\theta_2) > \frac{\tau p-(1-\tau)b}{\tau(b+b)}$ the taxpayer will strictly prefer to choose a lower σ since his expected payoff when lying exceeds the truth telling payoff. Since $G(x_\sigma|\theta_2)$ is increasing in σ and the possible values of $\frac{\tau p-(1-\tau)b}{\tau(b+b)}$ are in the interval $\left[0, \frac{p}{p+b}\right]$ this may occur for all τ and for all σ. It is *surely* the case if

$$G(x_0|\theta_2) > \frac{p}{p+b} \tag{12.11}$$

This situation occurs if the punishment is too small or the gain from getting away undetected is very big. The following result thus establishes a condition when the taxpayer always lies in equilibrium.

Lemma 12.3 *If the distribution $G(.|\theta_2)$ of the signals x in the case where the taxpayer is of type θ_2 is such that $G(x_0|\theta_2) > \frac{p}{p+b}$ then in every equilibrium the taxpayer lies with probability one and therefore chooses $\sigma = 0$.*

Proof: See Appendix.

It is interesting to note that from the perspective of the IRS – given an auditing technology described by $g(.|\theta_i)$ – for reasons which are entirely independent from any auditor-incentive scheme, auditing can never be effective. If the IRS faces a situation described by equation (12.11) any incentive scheme in the world will have no effect on enforcement. In this case only changes in the tax code (thus in b) the legal sanctions (i.e. p) can help, or the auditing technology must be improved, or the auditor and the taxpayers are motivated by other goals than direct monetary incentives.

Though there can be equilibria where the taxpayer always lies it is obvious that the given structure of the game does not allow for equilibria where the taxpayer tells the truth with probability one. This fact is summarised in the following:

Lemma 12.4 *There doesn't exist an equilibrium where the taxpayer reports his correct income with probability one.*

This is a standard result of auditing models without commitment possibilities. If audits have to be conducted by a self-interested auditor lack of commitment does not even have to be assumed but is implied by the structure of the problem. To see this (and prove Lemma 12.4) assume the taxpayer tells the truth with probability one. Then the self-interested auditor who has to move after he has received the report will always shirk. Due to the informational problem of unobservable effort post-contractual opportunism can never be prevented and so a commitment solution can't function with a self-interested auditor.

Equilibrium outcomes: incentive traps and imperfect but effective deterrence

Using these results we are ready to characterise possible equilibria depending on the exogenous parameters. Our approach is to partition the parameter space of our

model and pin down a unique equilibrium for different regions in this space. In this way we get a characterisation of possible equilibrium outcomes of our game. The equilibria broadly fall into two categories. In the first category the incentive scheme either doesn't work at all, or though it motivates the auditor to investigate reports properly, the deterrence effect is too weak to be effective. The second category pins down the case where deterrence works, though it can never work perfectly.

Incentive traps

The first class of equilibria that can occur are all characterised by the fact, that non-compliance will not be deterred. One obvious candidate for such a situation is the case where the auditor is paid so poorly that it can never pay for him to expend any effort. But there are also equilibria, where the auditor is a very alert investigator but still he is ineffective. We can summarise these cases compactly in the following proposition.

Proposition 12.1 *If*

$$I(0,x_0|L) < 0 \quad \text{or } G(x_0|\theta_2) \geq \frac{p}{p+b}$$

Or if simultaneously

$$G(x_0|\theta_2) < \frac{p}{p+b} \text{and } I(0,x_0|L) = 0$$

The only possible equilibrium outcome is one where the taxpayer always lies.

Proof: See Appendix.

The cases covered by this proposition are: (i) the trivial case where the incentives scheme is so unattractive that it can't motivate the auditor whatever the taxpayer might report; and (ii) the case where incentives work for motivation the auditor but not for deterring the taxpayer from underreporting. This shows that the idea of rewarding an auditor for detected lies for incentive reasons looked at in isolation is simple-minded and fails to consider the full strategic situation of the enforcement problem. As already mentioned in the comment to Lemma 12.3, we can see that the incentive scheme can work to motivate the auditor not to shirk but it can nevertheless be completely ineffective. The formula conveys when such a situation can occur: when punishments are low and the reward from evasion is high or the auditing technology is functioning poorly. The penalty p and the gain from underreporting b are determined by the law and can therefore not be chosen by the IRS. Thus even when it could choose an incentive wage scheme that motivates the auditor to investigate cases thoroughly it can be caught in a kind of trap. For incentive reasons it might want to induce the auditor to be interested that the taxpayer underreports because detection of such behaviour is the thing that pays off most for him. The idea of course is that this interest of the auditor will

prevent the taxpayer from lying *in equilibrium*. It might well be the case that the auditor investigates properly but the deterrence effect is too low to prevent misrepresentations of information by the taxpayer at all. In such a situation no incentive scheme in the world can make a self-interested auditor effective in enforcement. The analysis of the game thus clearly convey the interdependence of the auditor's and the taxpayer's incentives. From the perspective of the IRS the tax code and the legal system are given. All it can influence are the auditor's incentive scheme, i.e. the payoff r and f and possibly the auditing technology. Thus providing incentives by simply choosing the auditor payoffs, may lead to an expensive and ineffective enforcement process.

Imperfect but effective deterrence

Equilibria where the incentive scheme is able to make the auditor fulfil his deterrence function are described in the following proposition:

Proposition 12.2 *Let* $G(x_0|\theta_2) < \frac{p}{p+b}$ *and* $I(0,x_0|L) > 0$.

(i) If there exists a σ^* *such that* $G(x_{\sigma^*}|\theta_2) = \frac{p}{p+b}$ *and* $I(\sigma^*,x_{\sigma^*}|L) \geq 0$ *together with* $\tau = 1$ *this is the unique equilibrium.*

(ii) If σ^* *doesn't exist or implies* $I(\sigma^*,x_{\sigma^*}|L) < 0$ *there exists* $0 < \sigma < \sigma^*$ *such that* $I(\sigma,x_\sigma|L) = 0$ *and* $0 < \tau^* < 1$ *such that* $G(x_\sigma|\theta_2) = \frac{\tau p-(1-\tau)b}{\tau(b+p)}$, *and these equilibrium values are uniquely determined.*

Proof: See Appendix.

The proposition shows that punishment for an evader has to be sufficiently large to support an equilibrium where the incentive scheme for the auditor really works for improving enforcement. We need that $I(\sigma,\tau|\sigma = 0)$ is large or $\frac{p}{1+p}$ is large, i.e., p is large. If in addition the incentive scheme rewards detection of evaders sufficiently the auditor always checks suspicious reports. It can be seen from the proposition that it is optimal for the auditor to make his optimal 'false alarm' probability dependent on the lying strategy of the taxpayer and not only on the trade off between non-detection and 'false alarm' probability. The taxpayer is using a mixed strategy, which means that he doesn't underreport with probability one.

In the second case described in Proposition 12.2 the auditor is just indifferent between conducting a proper investigation or not. Thus the auditor shirks sometimes. Proposition 12.2 shows that the incentives of the auditor and the taxpayer are related in a rather complicated way. On the one hand, the optimal strategy of an auditor who has put in effort is determined by x_{σ^*}, which depends on the ratios $\frac{H(\theta_2|x,\sigma)}{H(\theta_1|x,\sigma)}$ and $\frac{f-e}{r+e}$. On the other hand the taxpayer's incentives depend on the ratio $G(x_{\sigma^*}|\theta_2) = \frac{p}{p+b}$, which in turn feeds back on the auditor's incentive to work properly.

The interdependence of auditor and taxpayer incentives

Our chapter is a first step into the analysis of auditor moral hazard and enforcement. We think, however, that one important aspect of the problem is clearly conveyed by our stylised model: the interdependence of the auditor's and the taxpayer's incentives. Since from the perspective of the IRS important parameters of the problem are already given, attempts to motivate the auditor by a straightforward incentive scheme, which makes him a 'bounty hunter' is naive and can lead to a poor and expensive performance by the entire enforcement process. This problem may be exacerbated by the fact that an incentive scheme, as the one proposed here, creates high stakes of collusion. We have assumed issues of collusion away in the present analysis for simplicity. However, the work of Adams (1993) shows by numerous examples from history that inviting collusion is a by product of 'bounty-hunter' incentive schemes. Our simple model shows how the interrelation between the auditor's and the taxpayer's incentives imposes limits on the resolution of auditor moral hazard by incentive contracts. However, our analysis also suggests some (moderately) speculative thoughts on the economics of enforcement in the presence of auditor moral hazard, which point in an interesting direction for future research on this topic.

So far we have interpreted the payoffs in the game strictly as expected utilities of money. The mainstream of economics takes it for granted that human behaviour is – more or less – entirely motivated by extrinsic monetary incentives. Though *terra incognita* to economists, it is well known to social psychologists and sociologists that so-called *intrinsic motivation* is another important determinant of human behaviour. Intrinsic motivation is related to activities an individual undertakes because 'one likes to do them or because the individual derives some satisfaction from doing his or her duty' (Frey and Oberholzer-Gee 1997: 746). Real world incentives usually consist of a complex interaction between these two forms of motivation. The importance of this aspect for economic analysis was repeatedly pointed out by Frey and others in a series of papers (see, for instance, Frey and Oberholzer-Gee, 1997; Frey and Holler, 1998). Also recent advances in experimental economics have repeatedly found evidence for intrinsic motivation in the form of reciprocity. Work by Fehr *et al.* 1997 provides experimental evidence for the importance of reciprocity in contract enforcement.

In the light of this literature our analysis highlights two things. First, we see that the tools of extrinsic motivation for the auditor are clearly limited and it might be worthwhile to think how intrinsic motivation may be employed to motivate otherwise self-interested auditors. But – and this is the second point – the issue is more intricate. The auditor in its very function is an instrument of extrinsic motivation of taxpayers through a process of monitoring. Work by Frey and Holler (1998) has pointed out the importance of the negative effect of excessive deterrence on intrinsic motivation to pay taxes. So a certain balance of intrinsic and extrinsic motivation also has an important role to play on the side of taxpayers and they are not independent of which incentives are there for the auditor. We think that our model might serve as a useful building block for a more systematic investigation

of this aspect. We believe that this is an interesting direction of future research on the economics of enforcement suggested by our work.

Conclusions

In this chapter we have discussed some consequences of auditor moral hazard on the economics of income tax policy enforcement. Providing a richer structure than traditional models of interactive tax policy enforcement, we analyse the interdependence of the auditor's and the taxpayer's incentives and discuss some implications of an imperfect auditing technology for tax enforcement. Our analysis points out major problems with such a solution: First, the penalties that can be applied to evaders as well as the cost for convicting evaders by a legal procedure are given for the IRS, it is limited in its possibilities of shaping incentives by only choosing a payment schedule for the auditor. It can be the case that any incentive scheme that rewards the auditor for correctly finding out evaders must fail. Second, the analysis suggests that the case of income tax enforcement and the economic resolution of auditor moral hazard might be an important case for an enforcement procedure that relies more on mechanisms of intrinsic motivation of auditors as well as taxpayers. Such an alternative approach to the analysis of enforcement problems would be an interesting direction for future research.

Appendix: proofs

Proof of Lemma 12.1: First we show that $I(\sigma,x_\sigma|L)$ is strictly decreasing in σ whenever it is positive. Suppose we had $q<r$ such that $I(r,x_r|L)\geq I(q,x_q|L)$. But since $E(x_r)<0$ we have $F(x_r)>0$. $q<r$ implies $1-q>1-r$ and so $(1-q)F(x_r)>(1-r)F(x_r)$ which implies

$$\begin{aligned}&\frac{\lambda}{\lambda+(1-\lambda)(1-q)}E(x_r)+\frac{(1-\lambda)(1-q)}{\lambda+(1-\lambda)(1-q)}F(x_r)\\&>\frac{\lambda}{\lambda+(1-\lambda)(1-r)}E(x_r)+\frac{(1-\lambda)(1-r)}{\lambda+(1-\lambda)(1-r)}\\&\geq\frac{\lambda}{\lambda+(1-\lambda)(1-q)}E(x_q)+\frac{(1-\lambda)(1-q)}{\lambda+(1-\lambda)(1-q)}F(x_q)\end{aligned}$$

Which is a contradiction to the fact that x_q were chosen optimally. Analogously we prove that if $I(\sigma,x_\sigma)=0$ (respectively $I(\sigma,x_\sigma|L)<0$) then for all $\sigma<r$ we have $I(r,x_r|L)<0$.

Proof of Lemma 12.3: If $G(x_0|\theta_2)>\frac{p}{p+b}$ then the expected payoff of lying strictly exceeds the expected payoff of truthful reporting even if $\tau=1$. $G(x_0|\theta_2)b+(1-G(x_0|\theta_2))(-p)>0$ if and only if $G(x_0|\theta_2)>\frac{p}{p+b}$. There can be no equilibrium τ,σ,x_σ with $\sigma>0$. Assume this is the case then the agent must be indifferent between lying and truth telling. Therefore

$$(1-\tau)b+(1-\sigma)\tau(G(x_\sigma|\theta_2)b+(1-G(x_\sigma|\theta_2))(-p)=0$$

$$\text{iff } G(x_\sigma|\theta_2) > \frac{\tau p-(1-\tau)b}{\tau(b+p)}$$

But this is impossible, since $G(x_\sigma|\theta_2) \geq G(x_0|\theta_2) > \frac{p}{p+b} \geq \frac{\tau p-(1-\tau)b}{\tau(b+p)}$ for all $0 \leq \tau \leq 1$ this is never possible.

Proof of Proposition 12.1: We describe all possible equilibria in which the agent always lies, i.e. when he chooses $\sigma = 0$. To consider the first case, if $I(0,x_0|L) < 0$ the auditor will never audit and therefore will choose $\tau = 0$. The agent will then always lie, whatever the actual value of $G(x_0|\theta_2)$. This follows form Lemma 12.1 because in this case $I(\sigma,x_\sigma|L) < 0$ for all σ. The second occurs when $G(x_0|\theta_2) > \frac{p}{p+b}$ then the agent always lies whatever the auditor's strategy. The auditor chooses $\tau = 1$ if $I(\sigma,x_\sigma|L) > 0$. If $I(\sigma,x_\sigma|L) = 0$ he chooses a $\tau \in [0,1]$. This claim follows from the fact that $G(x_\sigma|\theta_2) > \frac{p}{p+b}$ is increasing in σ for all σ. The third case occurs when $G(x_0|\theta_2) = \frac{p}{p+b}$. Again the auditor chooses $\tau = 1$ if $I(\sigma,x_\sigma|L) > 0$ and if $I(\sigma,x_\sigma|L) = 0$ he chooses a $\tau \in [0,1]$. In these cases there is no equilibrium with $\sigma > 0$ i.e. where the agent does not lie with probability 1. This is the case because the agent is indifferent between lying and telling the truth only if $G(x_\sigma|\theta_2) = \frac{\tau p-(1-\tau)b}{\tau(b+p)}$. Since $G(x_\sigma|\theta_2) > G(x_0|\theta_2)$ this would imply that the auditor always clears contradicting the assumption that $I(\sigma,x_\sigma|L) > 0$. To consider the last case. Assume that $G(x_0|\theta_2) < \frac{p}{p+b}$. Only if $I(\sigma,x_0|L) + 0$ the auditor is indifferent between conducting an inspection or not and therefore is willing to choose a τ such that $G(x_0|\theta_2) \geq \frac{\tau p-(1-\tau)b}{\tau(b+p)}$. These cases exhaust the possibilities described in Proposition 12.1.

Proof of Proposition 12.2: If there exists a σ^* such that $G(x_{\sigma^*}|\theta_2) = \frac{p}{p+b}$ and $I(\sigma^*,x_{\sigma^*}|L) \geq 0$ it is clear that we have a unique equilibrium with $\tau = 1$. The equilibrium is unique because for any $\sigma < \sigma^*$ we have $I(\sigma,x_\sigma|L) > I(\sigma^*,x^*|L) \geq 0$ and therefore the auditor would choose $\tau = 1$. But then $G(x_\sigma|\theta_2) < G(x_{\sigma^*}|\theta_2) = \frac{p}{p+b}$ which violates the equilibrium condition for the agent. For an analogous reason there doesn't exist an equilibrium with $\sigma > \sigma^*$. If such a σ^* doesn't exist or $I(\sigma^*,x_\sigma|L) < 0$ then there exists $0 < \sigma^* < \sigma$ such that $I(\sigma^*,x_{\sigma^*}|L) = 0$. Then the inspector is indifferent between all $\tau \in [0,1]$ and in equilibrium he chooses a τ^* such that $G(x_{\sigma^*}|\theta_2) \geq \frac{\tau^* p-(1-\tau^*)b}{\tau^*(b+p)}$. Any other equilibrium is impossible since as long as $I(\sigma,x_\sigma|L)$ is positive it is strictly increasing in σ.

Acknowledgements

We thank Bruno S. Frey, Simon Gächter, Michael Grosser, David Kelsey, Georg Kirchsteiger, Dilip Mookherjee, Manfred Nermuth, Gerhard Orosel, Stefan Reichelstein and seminar and conference participants at Alicante, Birmingham, St. Petersburg, Vienna and Warwick for helpful discussions and comments at various stages of this work. The usual disclaimer applies.

Notes

1 It must be noted, however, that a series of papers have pointed out the relevance of intrinsic motivation for economic analysis. Tax compliance is an example frequently used in this literature (see Frey and Holler 1997, 1998). In the more general context of constitutional design Frey (1997) has pointed out the importance of the complex feedback between intrinsic and extrinsic motivation.
2 As this is a first step we abstract from issues of collusion. Of course for incentive schemes as the one studied in this paper, collusion is a particularly relevant issue as documented in Adams (1993) by a number of historical examples.
3 Empirically also overpayment of taxes is observed. Andreoni *et al.* (1998) write that the taxpayer compliance measurement program reports for 1988 that 7 per cent of US households overpaid. Among those most overpaid only a relatively small amount. He will clear them without any effort. We can therefore concentrate on the part of the game tree where the auditor is confronted with a low-income report.

References

Adams, C. (1993) *For Good and Evil: The Impact of Taxes on the Course of Civilization*, Lanham, MD: Madison Books.
Andreoni, J., B. Erard and J. Feinstein (1998) 'Tax Compliance', *Journal of Economic Literature*, 34: 818–69.
Avenhaus, R., B. Von Stengel and S. Zamir (1995), *Inspection Games*, Discussion Paper no. 68, Jerusalem: Hebrew University.
Baiman, S., J.H. Evans and J. Noel (1987) 'Optimal contracts with a utility-maximizing auditor', *Journal of Accounting Research*, 25: 217–44.
Fehr, E., S. Gächter and G. Kirchsteiger (1997) 'Reciprocity as a contract enforcement device', *Econometrica*, 65: 833–60.
Frey, B. (1997) 'A constitution for knaves crowds out civic virtues', *Economic Journal*, 107: 1043–53.
Frey, B. and M. Holler (1997) 'Moral and immoral views on paying taxes', in S. Hellsten, M. Kopperi and O. Loukola (eds) *Taking the Liberal Challenge Seriously: Essays on Contemporary Liberalism at the Turn of the 21st Century*, Aldershot: Ashgate.
Frey, B and M. Holler (1998) 'Tax Compliance Policy Reconsidered', *Homo Oeconomicus* XV(1) Munich: ACCEDO Verlagsgesellschaft.
Frey, B. and F. Oberholzer-Gee (1997) 'The cost of price incentives: an empirical analysis of motivation crowding-out', *American Economic Review*, 87: 746–55.
Graetz, M., J. Reinganum and L. Wilde (1986) 'The tax compliance game: toward an interactive theory of law enforcement', *Journal of Law, Economics and Organization*, 2: 1–32.
Green, E. (1990) 'Classical statistics as a theory of incentives', *Games and Economic Behavior*, 2: 13–29.
Kofman, F. and J. Lawarree (1993) 'Collusion in hierarchical agencies', *Econometrica*, 61: 629–56.
Melumad, N. and D. Mookherjee (1989) 'Delegation as commitment: the case of income tax audits', *Rand Journal of Economics*, 20: 139–63.
Mookherjee, D. (1989) 'The economics of enforcement', in A.Bose, M.K. Rakshit and A. Sinha (eds) *Issues in Economic Theory and Public Policy: Essays in Honour of Tapas Majumdar*, New Delhi: Oxford University Press, 1997.
Reinganum, J. and L. Wilde (1986) 'Equilibrium verification and reporting policies in a model of tax compliance', *International Economic Review*, 27: 739–60.
Sanchez, I. and J. Sobel (1993) 'Hierarchical design and enforcement of income tax policies', *Journal of Public Economics*, 50: 345–69.

Shibano, T. (1990) 'Assessing audit risk from errors and irregularities', *Journal of Accounting Research*, 28: 110–47.

Index